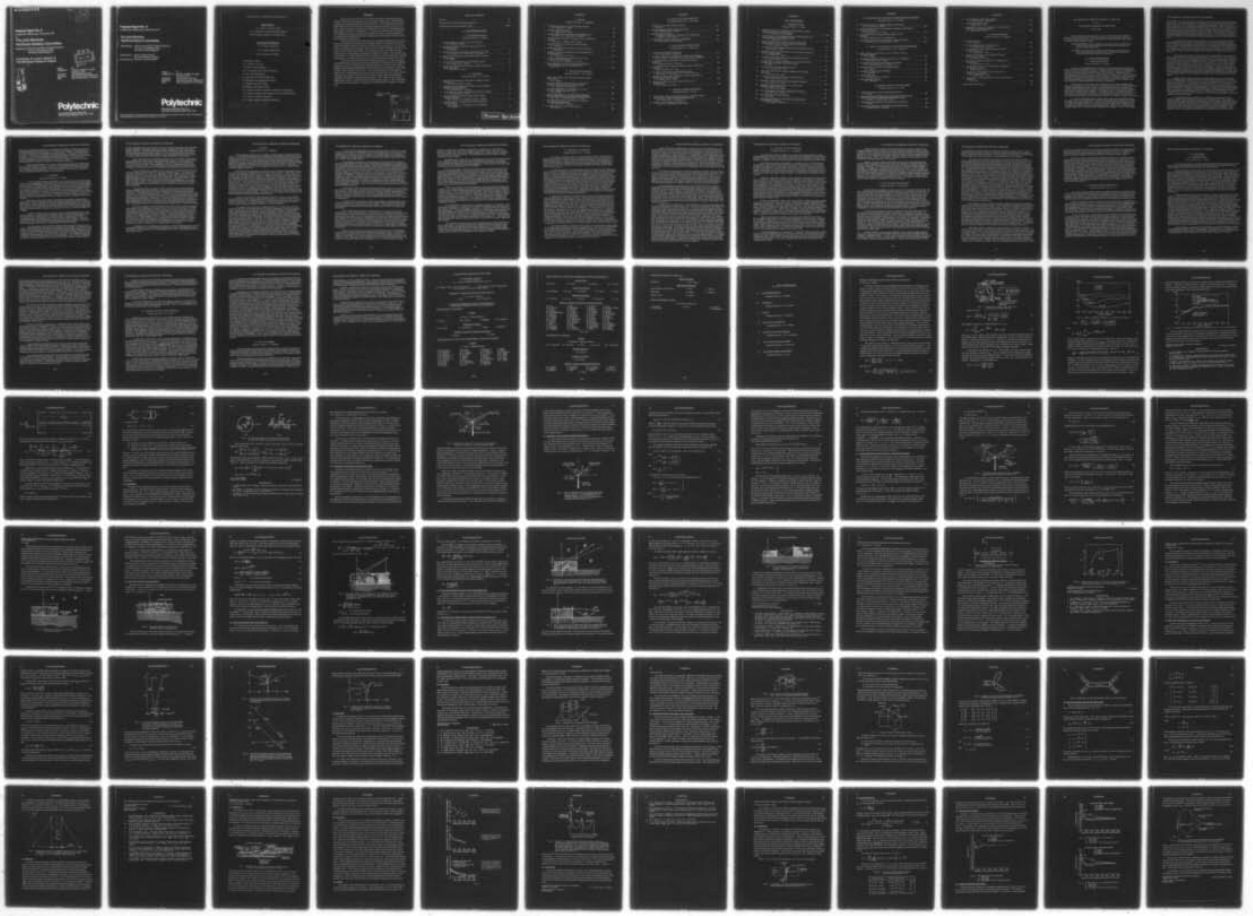
AD-A050 265 POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 20/3
PROGRESS REPORT NO. 42, 15 SEPTEMBER 1976 - 14 SEPTEMBER 1977, --ETC(U)
NOV 77 A A OLINER F44620-74-C-0056
UNCLASSIFIED POLY-MRI-452.42-77 NL

1 OF 7
AD
A050 265



AD A050265

(12)

Progress Report No. 42

15 September 1976 through 14 September 1977

to

The Joint Services Technical Advisory Committee

Representing: The Air Force Office of Scientific Research
The U.S. Army Research Office
The Office of Naval Research

A summary of current research at The Microwave Research Institute

DDC
RECEIVED
FEB 15 1978
F

AD No.
DDC FILE COPY

Report
R-452.42-77

Submitted
November
1977

for
Contract F44620-74-C-0056
Project No. 4751
This document has been approved for public release and sale; its distribution is unlimited.

Polytechnic

Polytechnic Institute of New York
333 Jay Street, Brooklyn, New York 11201

Progress Report No. 42

15 September 1976 through 14 September 1977

The Joint Services Technical Advisory Committee

Representing: The Air Force Office of Scientific Research
The U.S. Army Research Office
The Office of Naval Research

Submitted by: Arthur A. Oliner, Director
Microwave Research Institute
Coordinated by: Jerome Fox and Jean Maher

Report R-452.42-77

Submitted November 1977

for Contract F44620-74-C-0056
Project No. 4751

This document has been approved for public release and sale; its distribution is unlimited.

Polytechnic

Polytechnic Institute of New York
333 Jay Street, Brooklyn, New York 11201

Qualified requestors may obtain additional copies from the Defense Documentation Center; all others should apply to the Clearinghouse for Federal Scientific and Technical Information

MICROWAVE RESEARCH INSTITUTE

DIRECTORATE

A. A. Oliner: Director, MRI

J. Fox: Operations and Government Relations

S. W. Rosenthal: Budgets and Laboratories

MRI RESEARCH COUNCIL

A. A. Oliner, Chairman

J. Fox, Executive Secretary

S. W. Rosenthal, Ex Officio

L. Bergstein: Optics

H. L. Bertoni: Acoustics

A. *Hessel: Electromagnetics*

H. J. Juretschke: Solid State and Materials

L. Kurz: Communications

E. Levi: Electric Power Engineering

N. Marcuvitz: Wave-Matter Interactions

A. Papoulis: Data Processing

B. Senitzky: Solid State Devices

M. L. Shooman: Safety, Reliability and Software Engineering

E. J. Smith: Computers and Computer-Communication Networks

W. T. Walter: Quantum Electronics

D. C. Youla: Systems, Control and Networks

PREFACE

This report to the Joint Services Technical Advisory Committee presents a summary of the research programs in the broad field of electronics conducted during the past year at the Polytechnic Institute of New York. These programs are pursued within the framework of the Microwave Research Institute, and they involve the academic research activities of faculty in the departments of Electrical Engineering and Electrophysics, Physics, and Chemistry. The research projects cover a broad spectrum ranging from basic theoretical investigations in physics, applied mathematics, and engineering to experimental efforts involving basic measurements and the development of devices and materials.

The format of this annual report permits a coherent presentation of the various phases of the Joint Services Electronics Program (JSEP) at the Polytechnic and their relation to ongoing research in electronics sponsored by other agencies. This presentation is intended for the information of the Air Force Office of Scientific Research, the Army Research Office and the Office of Naval Research and, in addition, the other sponsors who are individually acknowledged throughout the report. The principal aims of the JSEP are to initiate deserving lines of research in a timely fashion and to develop investigations to a stature sufficient to attract individual support on their own merits.

In the early days of the Microwave Research Institute, the research program consisted primarily of projects involving electromagnetics and microwave components. Although the name of the Institute has remained the same, the nature of the research programs has broadened substantially, and the programs now encompass a wide range of topics within the field of electronics. The current programs are organized into twelve areas: electromagnetics; acoustics; optics; quantum electronics; solid state and materials; wave-matter interactions; electric power engineering; communications; computers and computer-communication networks; safety, reliability and software engineering; systems, control and networks; and data processing. A short description of the nature of these programs is presented in the Introduction, on pages xi through xxxii.

Arthur A. Oliner
Director

ACCESSION for
NTIS White Section
DDC Buff Section
UNANNOUNCED
JUSTIFICATION
BY
DISTRIBUTION/AVAILABILITY CODES
Dist. AVAIL. and/or SPECIAL
A

TABLE OF CONTENTS

Preface iii

The Microwave Research Institute Programs xi

Participating Faculty and Research Staff xxxiii

I. ELECTROPHYSICS

A. ELECTROMAGNETICS

Program Director: A. Hessel

1. Mutual Coupling in Arrays on Concave Surfaces
H. Ahn and A. Hessel 2

2. Cavity Resonances and Mutual Coupling in Concave Waveguide Arrays
H. Steyskal 6

3. Shadowing of an Inhomogeneous Plane Wave by an Edge
H.L. Bertoni, A.C. Green and L.B. Felsen 11

4. Radio-Wave Propagation Along Mixed Paths in Forests Environments
T. Tamir 22

5. Dielectric Grating Antennas for Millimeter Waves
S.T. Peng and A.A. Oliner 30

6. Novel Total-Absorption Phenomenon of Energy Incident on Planar Structures
V. Shah and T. Tamir 33

B. ACOUSTICS

Program Director: H.L. Bertoni

1. Shear Wave Propagation in Symmetric Hexagonal Honeycomb Cores
H.L. Bertoni and S.K. Park 41

2. Observation of the Leaky-Wave Field in an Acoustically-Excited Layered Structure
D.A. Davids and J. O'Brien 52

3. Leaky SAW and Guided Light Bragg Interaction in $LiNbO_3$
F. Freyre and W-C. Wang 57

4. Optical Imaging and Memory Using a Silicon Diode Storage Correlator
W-C. Wang, L.S. Rosenheck and K.C. Whang 63

5. Charge Storage in a Diode Mosaic Induced by Surface Acoustic Waves
W-C. Wang, L.S. Rosenheck, K.C. Whang and H. Schachter 70

Preceding Page BLANK

CONTENTS

C. OPTICS

Program Director: L. Bergstein

1. Perturbation Analysis for Optical Grating Coupler
   S. T. Peng and T. Tamir 78
2. Design of Grating Couplers
   T. Tamir and S. T. Peng 85
3. Network Methods for Dielectric-Grating Applications
   T. Tamir, S. T. Peng and K. C. Chang 90
4. On Simultaneous Blazing of Triangular Groove Diffraction Gratings
   L. S. Cheo, J. Shmoys and A. Hessel 96
5. Radiation Patterns of Dielectric Slab Waveguides
   L. Bergstein 102
6. Polarization Independent Dielectric Constant Profiles
   L. Bergstein 111
7. Leakage and Resonance Effects on Strip Waveguides for Integrated Optics
   S. T. Peng and A. A. Oliner 117
8. A Novel Leaky-Wave Strip Waveguide Directional Coupler for Integrated Optics
   E-W. Hu, S. T. Peng and A. A. Oliner 122
9. Photolithographic Lift Off of rf-Sputtered Dielectric Films With Application to Integrated Optics
   E-W. Hu 126

D. QUANTUM ELECTRONICS

Program Director: W. T. Walter

1. Superradiant Output from a Discharge-Heated Copper Vapor Laser
   W. T. Walter 133
2. Computer Modeling and Examination of Excitation Rates for a Bismuth Vapor Laser at 4722 Å
   W. T. Walter and N. Solimene 137
3. A Possible Explanation for the Absence of Laser Action in Bismuth Vapor at 4722 Å
   W. T. Walter 145
4. Rational Approximation of Temperature-Dependent Kinetic Coefficients for Efficient Computer Modeling
   G. M. Kull, N. Solimene and W. T. Walter 152
5. The Characterization of Noise Sources by Angular Correlation Measurements
   D. B. Scarl and S. McAfee 156
6. A Challenge for Non-Linear Spectroscopy
   S. R. Barone 158

CONTENTS

E. SOLID STATE AND MATERIALS

Program Director: H.J. Juretschke

1. A Solution of the X-Ray "Phase Problem"
   B. Post 163
2. Strain Dependence of Rust Layers at X-Band
   H.L. Bertoni and L.M. Silber 170
3. New Solid State Materials
   E. Banks, C. Bush, M. Greenblatt, B.R. McGarvey, S. Nakajima and M. Shone 178
4. Large Signal Transient Response of MOSFETS
   R.T. Kinasewitz and B. Senitzky 182

F. WAVE-MATTER INTERACTIONS

Program Director: N. Marcuvitz

1. Turbulence in Plasma-like Systems
   S.R. Barone and N. Marcuvitz 185
2. High Power Microwave Propagation Through the Atmosphere
   S.R. Barone, N. Marcuvitz, R. Pascone and N. Solimene 191
3. A Phenomenological Description of Intensity Dependent Reflection
   N. Solimene and M.C. Newstein 199
4. A Theoretical Study of Injection Tuning of Optical Parametric Oscillators
   E.S. Cassedy and M. Jain 205
5. Parametric Excitation of Coupled Plasma Waves
   B.R. Cheo, C. Hechtman, S.P. Kuo and T.Q. Yip 212
6. Evolution of Transverse Instability in a Hollow Cylindrical Weakly Ionized Plasma Column
   H. Kudyan and K. Chung 218
7. Response of Minor Species to Gravity Waves in the Thermosphere
   S.H. Gross and H. Eun 222

G. ELECTRIC POWER ENGINEERING

Program Director: E. Levi

1. New Results in Electrical Machine Theory: A Floquet-Theory of the General, Linear Non-Ideal Rotating Machine
   D.C. Youla and J.J. Bongiorno, Jr. 229
2. Design of Inductor Synchronous Motors
   E. Levi, L. Birenbaum and Z. Zabar 240

CONTENTS

II. SYSTEMS

A. COMMUNICATIONS

Program Director: L. Kurz

1. Maximum A Posteriori Estimator for Suppression of Interchannel Interference in FM Receivers
   F. A. Cassara and H. Schachter 248
2. A New Approach to the Edgeworth Series Approximation with Application to the Binomial Distribution
   C. Y. Chang and L. Kurz 253
3. Autocorrelation Function at the Output of Memoryless Nonlinearities Using a Differential Equation Approach
   R. Chassaing and L. Kurz 261
4. A Unique Frequency Division Multiplex Communication System
   L. Kurz and J. Reed 273
5. On Channel Monitoring Techniques in the Presence of Impulsive Noise or Fading Effects
   I. M. Habib and L. Kurz 275
6. Suppression of Timing Jitter in Partial Response Systems
   M. Kavehrad and L. Kurz 291
7. On Estimation of Channel Conditions in M-Ary Systems With Random Signals and Intersymbol Interference
   L. Kurz 296
8. A New Class of Non-rank Robust Detectors
   J. I. Cochrane and L. Kurz 301
9. The m-Interval Partition Detector (MIPD) with Dependent Input Data
   P. Kersten and L. Kurz 307
10. A Robustized Cumulative Decision Feedback System
    H. S. Ashtiani and L. Kurz 311
11. Sequential Partition Detectors with Dependent Sampling
    R. F. Dwyer and L. Kurz 319
12. Sequential Partition Detectors in Large Signal and Impulsive Noise Environments
    R. F. Dwyer and L. Kurz 333
13. Score Estimation for the Generalized Quantile Detector
    I. M. Habib and L. Kurz 345
14. Data Smoothing via Order Statistics
    S-G. Tyan 355
15. The Nyquist Problem in the Presence of Nonlinearities in the Duobinary System
    L. Kurz and M. Wernicki 358

CONTENTS

B. COMPUTERS AND COMPUTER-COMMUNICATION NETWORKS

Program Director: E. J. Smith

1. A New Approach to the Algebraic Theory of Automata
   T. T. Lee and E. J. Smith 364
2. A New Computer Hardware Description Language for the Study of Computer Structures
   M. Adamowicz and J. Mirza 375
3. Centralized Teleprocessing Network Design: A New Heuristic
   R. R. Boorstyn and A. Kershenbaum 390
4. Communication Networks with Adaptive Channel Capacities
   K. Sohraby, R. R. Boorstyn and L. Shaw 399

C. SAFETY, RELIABILITY AND SOFTWARE ENGINEERING

Program Director: M. L. Shooman

1. Upper Bounds on the Number of Tests Needed to Verify a Computer Program
   G. S. Popkin 408
2. Extensions of Software Physics to Measures of Complexity
   E. Berlinger and H. Ruston 412
3. Complexities of Natural and Computer Languages
   M. L. Shooman and A. E. Laemmel 414
4. Experimental Verification of Debugging Models
   D. L. Baggi 418
5. Automatic Programming Techniques
   E. Lipschitz 421
6. Small Scale Tests
   H. Ruston and M. L. Shooman 424
7. Micro Reliability Models
   M. L. Shooman 431

D. SYSTEMS, CONTROL AND NETWORKS

Program Director: D. C. Youla

1. Nonlinear Control of Linear Multivariable Systems
   L. Shaw 437
2. Optimal Replacement When Observable Stages of Deterioration Have Multivariate Exponentially Distributed Durations
   L. Shaw, C-L. Hsu and S-G. Tyan 442
3. Estimation and Modelling of Non-Stationary Time Series
   F. Kozin 444
4. Restoration of Impulse Train Signals from Smoothed Data
   A. Papoulis and C. Chamzas 452

CONTENTS

5. High Resolution in Pulse-Echo System
A. Papoulis and C. Chamzas 462

6. A Robustized Kalman Filter
I. Kadar and L. Kurz 471

7. Robustized Scalar Form of Gladyshev's Theorem with Application to Nonlinear Systems
I. Kadar and L. Kurz 488

E. IMAGE PROCESSING

Program Director: A. Papoulis

1. Signal Processing
A. Papoulis 501

2. Diffraction Elimination for Point Sources
A. Papoulis and C. Chamzas 503

3. A Contribution to Edge Detection Using Two-Way ANOVA Techniques
P. Legakis and L. Kurz 509

4. M-Gray Level Detection for Moderate and High Signal-to-Noise Ratios
P. Legakis and L. Kurz 515

5. Bivariate m-Interval Classifiers with Application to Edge Detection
P. Kersten and L. Kurz 524

6. A Robustized Vector Recursive Stabilizer Algorithm *for Image Restoration*
I. Kadar and L. Kurz 536

7. Estimation of Probability Density Functions Via Stochastic Approximation
P. Kersten and L. Kurz 547

8. Fixed Points of Running Medians and Locally Monotonic Regression
S-G. Tyan 551

Publications and Reports 561

THE MICROWAVE RESEARCH INSTITUTE PROGRAMS

of the

POLYTECHNIC INSTITUTE OF NEW YORK

(1976-1977)

This introductory section contains an over-all summary of the research programs in the broad field of electronics conducted at the Polytechnic Institute of New York during the past year. These programs are organized below under the following descriptive subject headings:

ELECTROPHYSICS: Electromagnetics; Acoustics; Optics; Quantum Electronics; Solid State and Materials; Wave-Matter Interactions; and Electric Power Engineering.

SYSTEMS: Communications; Computers and Computer-Communication Networks; Safety, Reliability and Software Engineering; Systems, Control and Networks; and Data Processing.

I. ELECTROPHYSICS

A. ELECTROMAGNETICS

Program Director: A. Hessel

The various investigations which fall under the broad heading of electromagnetics involve the propagation, guiding, radiation, and diffraction of electromagnetic waves in a large variety of environments. Included under this category during the past few years are major programs in wave types near interfaces, radiation from and scattering from periodic structures, various antenna investigations, particularly those involving planar and conformal phased arrays, radiation from sources and scattering by obstacles in media with relatively arbitrary properties, a systematic exploitation of ray-optical techniques in propagation, guiding, scattering and antenna problems, and topics related to biological hazards at microwave frequencies.

The program on wave types near interfaces some years ago introduced the concept of leaky waves, showed its value in the explanation of many radiation phenomena, and laid the foundations for a new class of traveling-wave antennas, the leaky-wave antenna, which permitted better agreement between theoretical and measured radiation patterns than any other type of antenna. The program led to a very general study of wave types, including several categories of complex guided wave and lateral wave, and showed their interrelations. The lateral wave was also shown to be the mechanism which permits most of the point-to-point communication in a jungle or forest environment.

A current investigation which relates to wave types near interfaces involves the lateral beam shift encountered by a beam of finite width when incident upon an interface under appropriate conditions. If the beam is incident upon the interface between two dielectric half-spaces at or very near to the critical angle of total reflection, the resulting beam shift is called the Goos-Hänchen shift. This shift is a special case of a larger class which is currently under systematic investigation. At the critical angle

of total reflection, the incident beam couples to a lateral wave which propagates along the interface. If a layer is present on the interface, or if a periodic grating is located there, and if the beam is incident at the angle corresponding to a leaky wave on the layer or grating, a very pronounced lateral beam shift is found. This beam shift, which is related to a pole in the wavenumber spectral plane, is much larger in general than the beam shift due to the lateral wave, which corresponds only to a branch point. This beam-shift mechanism has recently been used to achieve high coupling efficiencies at optical frequencies between laser beams and thin films. The beam shift itself offers a simple condition for optimizing the coupling parameters. The general properties of such beam shifts and their application to optical coupling devices are under systematic study.

The lateral beam shift is also required for a correct ray-optical model for modal propagation in thin films and optical fibers. Such modal propagation studies, which also involve the excitation of guiding structures of this type by Gaussian beams, and the optical coupling studies mentioned above, are considered in more detail in the Optics section. The leaky-wave beam shift is also shown to permit the optimization of the coupling efficiency of acoustic wedge transducers and couplers from acoustic surface waves to linear acoustic waveguides; these studies are reported in the Acoustics section.

The program on periodic structures has many ramifications. One phase relates to general studies of radiation from surfaces or interfaces which are modulated periodically, and in this context the Brillouin diagram for radiating periodic structures was first introduced and its usefulness clarified. This phase also included studies of various types of periodically-modulated slow-wave antennas. Another phase involved the study of special symmetries in periodic structures, particularly screw and glide symmetry. Studies of scattering by periodic surfaces led to a new and physically-satisfying theory of Wood's anomalies on optical gratings, an effect which remained poorly understood for 60 years. This theory shed substantial light on scattering resonances in general, and these ideas have been subsequently applied to other areas, such as scattering from plasmas and radiation from phased-array antennas. These studies are now being extended to the investigation of effects due to the boundedness of finite beams.

Another study relating to periodic structures led to a new way of treating mutual-coupling effects in phased arrays which takes these effects into account rigorously and automatically, and it proposed the first compensation scheme for the minimization of such effects. A recent study extended this technique to the explanation of resonance effects in large phased arrays, a phenomenon that can cause the array to become unexpectedly "blind" at certain scan angles. This theoretical and experimental study achieved a rather complete understanding of this phenomenon--its causes and how it may be avoided. In addition, a guided-wave approach has been applied specifically to the analysis of a slot-fed phased array on a conducting circular cylinder. This analysis is the first one for such a structure in which mutual-coupling effects have been completely taken into account.

A concentrated effort is under way to establish methods of analysis for mutually-coupled arrays of aperture elements on non-planar conducting convex and concave surfaces, using both modal and asymptotic (ray) methods. Non-planar arrays of this type are called "conformal arrays" because they are often located on the skins of aircraft, where they must "conform" to the shape of the structure (wing, fuselage, etc.) on which they are located. The motivation for this effort is twofold: a) for application to a recent interesting development in antenna arrays called the "Dome Antenna", which is a feed-array combination of a single planar phased array with a

passive, dome-shaped, feed-through lens, and b) for the design of flush-mounted slot arrays on conical or ogive surfaces for homing missile applications.

While the design of conformal arrays generally requires the knowledge of the coupling coefficients and element patterns in a mutually-coupled convex environment, a conformal feed-through scanning lens array design necessitates, in addition, the knowledge of mutual coupling for arrays on concave conducting surfaces. The following topics are currently being pursued: (1) the analysis of dually-polarized open-ended circular waveguide elements, and (2) the coupling coefficients in cylindrical arrays of waveguide-fed axial slits.

Analysis shows that the decay of the coupling coefficients in arrays on concave surfaces may differ radically from that on planar arrays. The decay rate may be considerably slower (even for $ka \sim 100$, where a is the local equivalent cylindrical radius). This feature indicates that the use of a planar array matching network may not be adequate, and that the size of the test array used for the measurement of coupling coefficients on concave surfaces should be larger than that used in the planar case.

The design of homing missile slot arrays requires the knowledge of coupling coefficients and element patterns on conical or ogive conducting surfaces. A computer program based on the harmonic series solution for a conical geometry is extremely time consuming. To alleviate this difficulty, analytical expressions based on principles of the Geometric Theory of Diffraction (GTD) have been obtained which include modifications to torsional geometry of the existing GTD formulas valid for the torsionless case. Based on these expressions, a computer program has been developed for the calculation of the GTD radial or circumferential slot far-field patterns on conical surfaces. The numerical results compare well with those based on the harmonic series, except in the forward axial (tip) region, and in regions of low intensity in the GTD expressions, where the tip scattering that was neglected in the GTD analysis is of importance. For the computation of mutual coupling, GTD expressions were derived for the surface current due to a magnetic current element located on a conical surface, based on the asymptotic solution of the canonical problem of a circular cylinder. For axial slots, the mutual admittance values calculated via the GTD expressions for a circular cylindrical surface for $ka \sim 10$ check in the deep shadow and for the near neighbors with harmonic series results found in the literature. For circumferential slots, the asymptotic expressions must include additional truncation functions.

The current program includes the following topics: the analysis of spiral elements for phased array applications; the performance of a two-dimensional scanned lens array; modal analyses of mutual coupling on concave spherical surfaces; ray methods for two- and three-dimensional concave surfaces; and the synthesis of wide-angle scanning lens arrays.

The program on propagation and scattering in media with arbitrary properties was motivated by the recognition that recent technology requires a knowledge of the propagation of electromagnetic waves in physical environments of increased complexity, and of the radiation characteristics of antennas and the scattering characteristics of obstacles in these environments. Examples of such media are provided by ionized plasmas, either in the laboratory or in outer space, whose electrical properties can be described macroscopically in terms of an inhomogeneous or homogeneous isotropic dielectric, an anisotropic dielectric, or a mechanically deformable dielectric material, the choice of a particular model being dependent on the circumstances in question. In addition, turbulent processes in the medium may require the inclusion of statistical properties in its description. When a plasma medium surrounds an antenna or a scattering object, a situation encountered when a rocket or satellite passes through the

ionosphere or when a high speed vehicle re-enters the upper atmosphere, the above-mentioned processes of radiation and scattering become relevant for problems of radio communication and detection. They are also relevant for optical detection of satellites and other objects since the optical signal traverses an inhomogeneous and turbulent atmosphere on its way to the ground-based detector.

In these studies, ray-optical concepts play an important role. The exploration of the range of validity of ray-optical procedures in propagation, guiding, radiation and scattering in a general environment, and the subsequent application of these techniques to problems of current interest, provide the basis for much of this quasi-optic phase of the electromagnetics program.

Particular attention in the quasi-optic program has been given to the following subject areas: 1) propagation of pulses in a dispersive environment; 2) scattering by discontinuities in homogeneously and inhomogeneously filled waveguides or ducts; 3) propagation and scattering of exponentially-decaying (evanescent) fields in lossless or lossy regions by complex ray techniques; 4) inhomogeneous wave tracking. To deal with the first category, an asymptotic theory based on space-time rays (a generalization of geometric-optical rays) and plane-wave dispersion surfaces has been developed. Complicated transient wave processes are thereby described in terms of the motion, evolution, and interaction of wave packets. The theory has been applied to spatially-varying, temporally-varying and moving plasma media, and the results obtained are relevant for the study of pulse degradation, pulse compression and related phenomena.

In the second category, novel applications of ray-optical (i.e., high-frequency) methods have been found to yield remarkably accurate results for radiation from, and scattering by, discontinuity elements in guiding structures, even in the relatively-low-frequency propagation range of only the dominant mode. The technique has recently been applied to VLF mode excitation and conversion in the earth-ionosphere waveguide, and has now been extended to multiwave media such as compressible plasmas or elastic solids. Moreover, by viewing stable and unstable optical resonators as open-ended waveguides along the coordinate transverse to the mirror axis, a novel and promising analytical method for determining losses and field configurations in open resonators is presently under study. The method has so far been remarkably successful in explaining, by a simple model, the complicated loss behavior of eigenmodes in unstable resonators, which are widely employed with high-power lasers. This study is discussed further in the Optics section.

Emphasis in the third category is placed on determining the local propagation properties of evanescent fields for the purpose of developing therefrom a theory of propagation and scattering analogous to the geometrical theory of diffraction for non-evanescent fields. While evanescent fields can be regarded as traveling along ray trajectories in complex space, the investigation seeks to clarify the physical significance of "complex rays" and their possible utility in constructing the fields. Results obtained are of interest for communication with, or detection of, objects located in the refraction shadow region of an antenna illuminating an inhomogeneous lossless environment, and for similar problems involving lossy media; for gap coupling to totally-reflected optical beams; and for calculation of diffraction losses in open optical resonators.

Alternative to the complex ray approach is a new theory, which provides a means of tracking inhomogeneous wave fields in real coordinate space. Examples of inhomogeneous wave fields are evanescent fields, leaky waves and Gaussian beams. The advantage of this theory is that it provides direct information on phase front distortion and amplitude changes as the wave or beam penetrates a medium and is reflected,

refracted or scattered; in particular, the theory accommodates scattering events that lead to substantial field distortions as, for example, scattering of a Gaussian beam by an obstacle whose size is comparable to the beam width. Results have been obtained for various evanescent wave and beam scattering problems, and further studies are in progress. A very recent development involves application of evanescent wave tracking to modal propagation in graded-index optical films and fibers. This entirely new analysis promises to render tractable a broader class of index variations than can be accommodated by presently used modal or ray-optical techniques.

It has long been known that there are various biological hazards attendant on the use of high microwave power. The human is particularly sensitive in this respect. There exists, therefore, strong motivation for the establishment of tolerance limits and, ultimately, safety standards. A study of the effects of microwave radiation on the eye, employing rabbits as subjects, is continuing.

B. ACOUSTICS

Program Director: H. L. Bertoni

A major program that has been underway for several years deals with guided acoustic waves propagating on the surfaces of, or at interfaces between, elastic solids. Because of the extremely low velocity of elastic waves, as compared to electromagnetic waves, and their correspondingly small wavelength, these waves have found application in a variety of miniaturized signal processing devices, whose electromagnetic counterparts would be cumbersome, or not at all feasible. Such devices include delay lines, pulse compression filters, band pass filters, high frequency resonators and, in conjunction with semiconductors, convolvers, correlators and storage elements. These devices involve acoustic wave interactions and are therefore acoustic in nature, but they are fitted with input and output electro-acoustic transducers and are employed in electronic systems.

Motivated by these applications, the program has sought to explore device applications as well as the basic wave scattering and coupling phenomena that occur in such devices. Many of these scattering and coupling phenomena also arise in methods which are used for the nondestructive testing of solid materials for flaws, cracks, etc. Our program is currently exploring the ways in which known methods of nondestructive testing or evaluation can be made more quantitatively accurate, and also devising new methods.

In order to study acoustic wave scattering and coupling phenomena in a systematic manner, we introduced and developed a microwave network and transmission line formalism for acoustic waves in isotropic solids. Using this formulation, transmission-line representations have been derived for bulk acoustic waves, and equivalent networks were obtained for several types of planar interface. These networks in turn were used to derive the characteristics of waves guided by a variety of planar structures, such as free or welded plates and plated surfaces.

The transmission line and network approach also forms the basis for a systematic investigation of the properties of several basic waveguides for acoustic surface waves. These waveguides are linear (as opposed to planar) configurations which serve to laterally confine the acoustic fields; the structures under detailed examination in the program include the strip and slot waveguides, which are flat overlay guides, and the rectangular ridge topographic and tall overlay waveguides. The studies are primarily theoretical, but some measurements are also being taken. The analytical results for the strip and slot waveguides are the most accurate ones available, and the analyses

for the rectangular ridge structures form the only analytical theories to date for those guides, although very accurate numerical results have been published for the topographic structure. These analytical expressions, furthermore, agree well with both our own measurements and those of others. We are also studying the properties of the ridge structure of semi-infinite height, which can guide waves along its top edge, in order to better understand the behavior of ridge waveguides of finite height.

The transmission line and network approach is not only useful in the analysis of waveguides, but it has also permitted quantitative design in a novel method for the efficient excitation of these linear waveguides. This method is a two-dimensional version of the prism coupler used in integrated optics for coupling a laser beam to a surface wave on a thin film. In its application to the excitation of the acoustic strip waveguide, an efficiency of 65% over a 25% band was achieved without any cut and try.

Novel components for acoustic surface waves are also being devised, and they are being analyzed with the help of the networks mentioned above. Among these components are simple filters, bulk wave suppressors, power splitters, nonreflecting delay-line taps, and resonant cavities, all based on either total reflection or partial reflection produced by a narrow strip of plated material. These components have the added advantage that they can be used on any substrate, piezoelectric or not. The effect of a beam of finite width on the performance of these components is presently under theoretical study. A transmission line representation for bulk waves in piezoelectric crystals has also been developed, and it was used to treat a novel class of bulk wave filters.

An important class of surface acoustic wave devices for signal processing employs nonlinear space charge effects in semiconductors adjacent to piezoelectric materials. Studies being performed here relating to this class of devices include convolvers, an acoustic phase-locked loop, correlators, time compressors and storage devices. The use of semiconductors in proximity to a piezoelectric substrate was first proposed at the Polytechnic; subsequent work here has added to both the theoretical and experimental development of such devices. Besides the original work on convolvers, the acoustic phase-locked loop was invented here, and a fully integrated version is now under development. Real time correlation and time compression were also obtained here, and these devices are currently being refined, together with a storage correlator.

The first complete theoretical analysis has been carried out of the phenomena associated with the reflection of a bounded acoustic beam from the surface of a solid at the angle for phase matching with the surface wave. The primary effect is a large lateral shift of the reflected beam from the position predicted by geometrical optics. Experimental observations carried out elsewhere for the case of a beam incident from a fluid are in complete agreement with the theory. Subsequent studies for a beam incident from a second solid have led to a complete theoretical understanding of the performance of wedge transducers for exciting surface waves; this theory has shown, for the first time, how high efficiencies can be achieved for such a device. These phenomena also offer a new method for the nondestructive evaluation of the surface properties of a solid, and their use is currently under investigation.

In addition, we are studying the propagation of waves resulting from the coupling of magnetic moments in a ferrite material to the elastic stress. The dispersion and polarization characteristics of such magnetoelastic waves have been found for both bulk and surface waves.

C. OPTICS

Program Director: L. Bergstein

Great progress has been made recently in broadening the research program in optics. The theoretical studies have been extended and the experimental effort has been expanded to include the fabrication and measurement of integrated optics structures and related thin-film devices. The present program consists of the following topics: the development of a new narrow-band matched optical filter (theoretical and experimental), research in integrated optics and related thin-film devices (theoretical and experimental), and the investigation of unstable optical resonators (theoretical only). The topic of integrated optics is the one which is expanding most rapidly, and it currently consists of several different projects.

Development is progressing on a class of narrow-band optical passband and/or rejection filters based on selective reflection from atomic vapors. These filters, which can operate in the 580 to 8550Å range, are rugged, of small size, and have a large acceptance angle and high spectral resolution. Because of these factors, the vapor filters are expected to be superior to presently available devices in pollution monitoring and other applications which require simultaneous multichannel spectrochemical analysis for trace elements. They are also expected to be useful in the detection of isotope concentrations, as a narrow-bandwidth ultraviolet mirror for high intensity radiation, and in astronomical observations. Tests performed on a mercury vapor filter verified the theoretical predictions. With a center wavelength of 2537Å, the filter has shown a peak reflectance of 40%, an acceptance angle of 7 degrees, and a bandwidth of only 0.14Å. The filter was used in an emission flame photometer to perform a spectrochemical analysis for mercury in a water solution. The device was also used to take a monochromatic photograph of the spatial distribution of excited mercury in the flame from which an estimate of the flame temperature profile could be made.

This device principle is being extended to elements which are more difficult to use since they require a higher operating temperature. An example of this type of element is sodium. If a successful filter can be built using sodium, the same structure can be used to contain a variety of elements. At present the fabrication of a cell utilizing a nickel housing and sapphire windows is nearing completion.

The purpose of the program on integrated optics is to develop film-layered structures and other micro-optic devices for eventual application in optical communications. It began with an investigation of the coupling and progression of optical beams along and through thin films and other optical waveguides. Planar and cylindrical polarization-independent waveguides have been investigated and were shown to be practical. Such waveguides can support non-polarized beams and can transmit information at twice the rate of ordinary waveguides. For planar waveguides it was found that phase and group velocity matching of a TE and a TM signal can be achieved by the addition of only a single thin-film layer. Such a scheme can also be used in nonlinear interaction applications to match the group velocity of two optical signals of different wavelengths and/or field distributions. Results analogous to those found for planar waveguides have also been found for cladded fibers. An approximate (but general) transmission line equivalent which we have developed for this class of structures reduces considerably the analytical complexity of the problem and allows the determination of the modal dispersion characteristics in a fairly straightforward manner. Moreover, we have found a realizable radially-variable refractive index distribution n(r) for which all rotationally-symmetrical TE and TM modes have the same group velocity and spatial power distribution. The modes and eigenvalues of such an optical fiber and its dispersion properties were found.

THE MICROWAVE RESEARCH INSTITUTE PROGRAMS

The beam-shift phenomenon which occurs when a beam is incident on a planar structure at the angle corresponding to a leaky wave supportable by the structure was mentioned in the section on Electromagnetics. This phenomenon was investigated theoretically in quite general terms, and the conditions for optimum beam-to-surface-wave coupling due to optical beams of finite width were derived for both uniform thin-film couplers and grating couplers.

The techniques developed by us previously for treating periodic structures have been combined with the beam-shift considerations and applied to the detailed analysis and design of planar grating couplers. These integrated-optics components serve the important function of coupling an optical beam into (or out of) a surface wave guided by thin films. In particular, an exact solution of the boundary-value problem posed by dielectric gratings has been formulated and a computer program has been set up to generate highly accurate numerical results. In this manner, the coupling efficiency, operating conditions and other data have been determined for a large variety of grating couplers. In addition, a simplified perturbation procedure requiring much shorter computation times has also been developed and found to yield excellent approximations of the exact results. These techniques are being used to determine dimensions for a class of blazed gratings that are capable of a substantially higher coupling efficiency than that of ordinary gratings. All of these techniques are instrumental in providing a range of simple criteria for the design of dielectric gratings having desirable diffraction and waveguiding properties.

A laboratory facility has been established and experiments are now in progress to verify some of the analytical results and to assess the performance of certain micro-optic devices. The facility includes a capability to produce solution-deposited films with a view towards making active devices, and a scanning electron microscope which has been modified to write patterns for optical waveguides, gratings and circuits. A computer-controlled digital interface was developed to generate and control the desired scanning pattern.

A model has been developed to describe the exposure and development characteristics of photoresist. The accuracy of this model has been verified and the model has been applied to the fabrication of grating couplers. The couplers so constructed were found to have properties that agree well with the predictions of the theory mentioned above.

Several additional theoretical investigations have recently been initiated relating to the area of integrated optics. One of these is an extension of the beam-shift considerations, mentioned in the Electromagnetics section, involving lateral waves (rather than leaky waves, which are appropriate to the grating couplers mentioned above). These lateral ray and beam shifts on boundaries with incidence-angle-dependent reflection coefficients are currently receiving special attention because it has been recognized that the correct ray-optical model for modal propagation in films and layers requires laterally shifted paths. A systematic study of lateral shifts, both longitudinal (Goos-Hänchen type) and transverse (Imbert type), for isotropic and anisotropic, homogeneous and graded-index, two-dimensional and three-dimensional, guiding regions is presently in progress.

Another new theoretical area involves the determination of the propagation characteristics of linear waveguiding structures for integrated optics. Under investigation are two waveguides which are composed of metal-clad regions on thin films located on a substrate. One of these guides consists of a slot between two metal-clad regions; a basic constituent mode was neglected in analyses performed elsewhere, and we find that important performance differences occur in a certain frequency range.

The other waveguide is a novel one consisting of a metal strip on a thin film on a substrate. This guide employs the "plasmon" mode and is lossier than other known linear waveguides, but it seems to be dispersionless over a large frequency range and may therefore be of distinct interest. The properties of these two guides are being examined in greater detail.

An additional theoretical topic is an application of the technique for the tracking of inhomogeneous fields, mentioned in the Electromagnetics section. As indicated there, application is being made to modal propagation in graded-index optical films and fibers; this new approach permits the treatment of a wider class of refractive index variations than can be handled by other techniques.

Several novel thin-film devices have been devised and investigated both theoretically and experimentally. Thin-film polarizers, beam splitters, and (fixed and tunable) narrow-bandwidth spectral filters for the visible and 10-micron region have been investigated and were shown to be feasible in both the usual optical beam-shaping applications and in integrated optics applications, where these signal shaping components can become part of the coupling scheme. Experimental tests have confirmed the theoretical results. A five-layer FTR narrow-bandwidth spectral filter and a single-layer FTR beamsplitter with transmission properties which are independent of the polarization of the incident beam were developed and tested and their properties were found to agree with the theoretical predictions.

Another new project involves the analysis of certain geometrical discontinuity problems. As a first example, we have solved an important discontinuity problem for a class of planar optical waveguides; the discontinuity consists of a change in height of a dielectric layer. In our modal expansion approach, which is fairly general, we place the open waveguide structure between two conducting planes and then allow the separation between the planes to reach infinity. This greatly reduces the complexity of the problem and permits the solution of a wider class of problems than is possible with other methods found in the literature. The solution we have found for single-mode waveguides agrees well with those found in the literature by less general and/or more cumbersome methods.

The investigation of the properties of unstable resonators is continuing. Two independent approaches are employed; one involves a longitudinal approach whereas the other uses a transverse, or waveguide, viewpoint, discussed briefly in the section on Electromagnetics. Using the longitudinal approach, a meaningful, approximate solution has been found for the complete set of modes and associated eigenvalues of optical resonators filled with an active medium having transverse variations in both the gain and refractive index. This has led to a simple physical picture which permits substantial new insight into the mechanism of operation of such resonators and will help in the determination of the optimum resonator parameters which are to be used in high-power laser devices.

For unstable resonators whose transverse boundaries are determined by the mirror dimensions, the waveguide approach to unstable resonators formed by hyperbolic strip mirrors has been found successful in explaining in a remarkably simple manner the complicated behavior of the complex eigenvalues of the resonant modes. These results are now being extended to circular mirror configurations and attention is being given to the effects of inhomogeneity in the resonator medium.

D. QUANTUM ELECTRONICS

Program Director: W. T. Walter

The research program in the area of Quantum Electronics comprises that portion of electronics in which quantum mechanical effects become important because of the discrete nature of charge carriers, matter or radiation. Optics and Wave-Matter Interactions are related research areas whose boundaries with Quantum Electronics are not distinct. Some of the research previously reported in the Quantum Electronics area has been shifted in this report to the Optics or Wave-Matter Interaction research areas.

A research program has been developed which has attacked outstanding problems of quantum electronics on a broad front. Significant results include: the development of a new type of optical and infrared spectroscopy which has demonstrated more than two orders of magnitude improvement in resolution compared to previous techniques, a corresponding improvement in the long-term frequency stabilization of lasers, significant progress in the understanding of the irreducible quantum fluctuations which are present in the output of lasers and other quantum electronic devices, advances in the understanding of the propagation of ultra-short pulses in resonant media, progress in the development of a new wide-angle low-noise laser receiver, an improved understanding of the properties of the acousto-optical interaction and the use of this interaction to provide acoustical gain and also to mode lock a ruby laser, a significant improvement in the brightness and in the average power generation density of the very-high-gain pulsed copper vapor laser, construction of a computer model to analyze these pulsed metal vapor lasers, and the development of a procedure for calculating improved values of atomic transition probabilities.

One group of projects has centered on the study of the nonlinear response of saturable absorbers to intense laser beams. These projects have yielded a new type of ultra-high-resolution spectroscopy, significant improvement in the long-term frequency stabilization of lasers, and have also contributed to the progress on an automatically-adapting filter which offers the promise of a significant improvement in the performance of an optical radar receiver. Laser Saturated Resonance Spectroscopy (LSRS) has been applied to the study of the absorption spectrum of the SF_6 molecule at 10.6μm and the I_2 and Br_2 molecules at 0.6328μm. A resolving power of 10^8 was demonstrated, representing a two order of magnitude improvement over conventional spectroscopic techniques. The study of the LSRS of xenon has now been completed. All of the observed components in the LSRS spectrum of a naturally occurring mixture of xenon isotopes (masses 128, 129, 130, 132, 134 and 136) at 3.5μm have been resolved and identified. The narrow Lamb dips have also been used to stabilize the frequencies of both the CO_2 (10.6 μm) and He-Ne (0.6328 μm) lasers. A laser which is frequency stabilized to a Lamb dip is a strong candidate for selection as an optical (or infrared) wavelength standard, and will also be useful for long-range interferometry, for deep-space communication links, for coherent optical radar, and for seismic studies.

The high intensities achievable in laser beams can perturb atomic energy levels and radically alter the position and shape of spectral lines. The theory of these phenomena is being extended and refined. Recent studies have centered on the inclusion of finite lifetime and recoil effects. New results were obtained on the effect on the molecular center-of-mass motion of a resonant driving field. The analysis exploited a novel calculation technique based on the application of Floquet's theorem to periodically-driven quantum-mechanical systems. It is hoped that these interesting effects at optical frequencies will eventually be observed and compared with theory.

Another theoretical investigation in progress aims at understanding the irreducible quantum fluctuations always present in the output of lasers and other quantum electronic devices, from the point of view of first principles. A study has been completed of the effect of an internal homogeneously-broadened saturable absorber on the intrinsic fluctuations of a laser oscillator. A related effort is concerned with the understanding of spontaneous emission processes in resonant dispersive configurations other than free space. This problem has been solved in general terms by relating it to the problem of calculating thermal field distributions. Recently, the spontaneous emission noise and the response characteristics of driven quantum electronic systems have been evaluated for classes of spatially degenerate molecules. Effects associated with the nonstationary character of the correlation functions have been particularly emphasized. These effects have been overlooked in recently published work on the response of driven systems.

The study of the propagation of ultra-short pulses in resonant media has led to the discovery of new relationships between their spatial and temporal structures. These include: strong self focusing, phase modulation which accompanies transverse confinement of the pulse, additional unsuspected structure of plane wave pulses, some effects implied by the inclusion of a finite transverse relaxation time, the transverse stability of confined fields, new types of distortionless fields which simultaneously possess phase modulation, polarization modulation and three-dimensional structure, and some effects produced by a nonlinear index of refraction of the host material.

The spectroscopic investigations of saturable absorbers and the study of propagation of ultra-short pulses in resonant media relate directly to the feasibility study of the low-noise wide-angle BALAD (Bleachable Absorber Laser Amplifier-Detector) receiver. In this receiver, an absorption cell (filled for example with SF_6) is placed between the laser amplifier and the detector to serve as a threshold filter, both in the frequency domain and in two spatial dimensions, and it screens out all background light which is not coherent with the signal beam. Calculations predict superior performance in optical radar and communications applications. The LSRS studies have clarified the effect on laser amplifier gain, bandwidth and band center that would result from the use of separated single xenon isotope samples in an optical radar system with a BALAD receiver.

Visible and near-visible lasers with an efficiency and power comparable to the far infrared CO_2 laser are still lacking. The vapors of certain metals and transition elements have attractive energy level structures. A number of high-temperature laser systems have been constructed to explore the most-promising metal vapor laser media. The pulsed copper vapor laser has high gain and high peak power at 5106Å. Moreover, unlike other self-terminating lasers, single mode output and therefore high peak brightness can be obtained. Recent experiments using higher pulse repetition rates and a larger volume have produced an average power of 11 watts. With argon as an additive gas, average powers above one watt have been achieved at 5106Å from an active region only 10 cm long and 1.25 cm in diameter, corresponding to an average power production of 0.1 watts per cubic centimeter. A heat-pipe copper vapor laser has been demonstrated, and preliminary experiments with a high-loss resonator have yielded a peak brightness of 10^{12} watts per square centimeter per steradian, which is the highest brightness measured for a gas laser. Heat-pipes and the use of dissipated discharge energy for self-heating have potential for practical sealed-off copper vapor lasers with average output powers up to 100 watts in the visible. Still higher output powers and overall electrical efficiencies of up to 10% are anticipated from the current investigation of impedance tailoring of copper vapor systems. Other metal vapor systems are being examined for efficient laser action in the ultraviolet and blue-green spectral regions. A computer model has been constructed and is being used in conjunction with the experimental analysis of several metal systems. Theoretical analysis of these systems for higher efficiency lasers has led to the development of a procedure for calculating improved values of atomic transition probabilities from spectral intensities.

E. SOLID STATE AND MATERIALS

Program Director: H. J. Juretschke

The solid state research ranges over a wide area of experimental activities encompassing the creation of new or better materials, the investigation of many specific properties of solids, and the application of these materials and effects to practical situations. In addition, there are active programs in various aspects of solid state theory.

In the area of materials development, an important sector is concerned with new magnetic compounds involving rare earth and transition metal ions, and with the production of phosphors for IR-to-visible conversion. Programs involving thin films include the development of high-mobility semiconducting layers, and the preparation of highest quality single crystal films of the noble and ferromagnetic metals and the semimetals. In addition, there are systematic efforts to produce good insulating barriers on semiconductor and metal single crystal surfaces, for application to tunneling and other devices.

Among the properties of interest, those related to electromagnetic interactions play a central role. Electron transport in semiconductors is studied for hot-electron nonlinear effects, and for surface interactions. Size and surface effects in carrier transport in magnetic fields, especially as related to the band structure and to surface relaxation mechanisms, are investigated in thin films. One group of studies centers on the spin dependence of electrical conduction in magnetic materials, while another is concerned with the characteristics of thin tunneling junctions, either vacuum or insulating, between normal or superconducting materials. New research includes the interaction of x-rays with electrons in semiconductors through the internal photoelectric effect, the determination of charge concentrations at metal-insulator interfaces by surface electron scattering, the details of metallic surface self diffusion as it affects the electrical surface properties, and the concentration profiles in alloys of thin films as affected by the large surface-to-volume ratio.

Studies in magnetism include a study of the magnetic coupling between transition metal ions in complex fluorides, using neutron diffraction, magnetic susceptibility and Mössbauer spectroscopy. Magnetic resonance measurements (NMR, ESR) are being used to study clustering of rare-earth ions doped in CdF_2. This material has been developed as a base for infrared to visible light conversion, which has potential application in a variety of display devices. Another study is looking at the origins of the frequency dependence of magnetic anisotropy in ferrites and at the anisotropy of their magnetogyric ratio. A program in magnetic resonance of ESR and NMR is elucidating details of electronic structure in many of the materials used in other investigations.

Optical investigations include a broad investigation of the Faraday effect in ferromagnetic metals that aims at clarifying details of their band structure, especially as a function of temperature. At very short wavelengths, in the x-ray region, studies are pursued on quantitative aspects of the resonant Borrmann transmission in perfect single crystals under conditions where three or more distinct beams are in strong interaction. Here an analysis of loss mechanisms has been critical, and we have initiated work on better determination of x-ray dispersion coefficients, and on the energy distribution of the propagating waves relative to the atomic sites.

X-ray investigations are concerned with crystal structures, perfection, thermal effects and charge distributions in a wide range of materials. Determination of lattice constants through multiple diffraction measurements are yielding data of considerably enhanced accuracy and precision. Multiple diffraction effects are also being applied in a careful interpretation of x-ray data, including linear absorption coefficients, line shapes and line shifts.

Theoretical studies include a fundamental program in electron-photon interactions in magnetic fields, and the theory of slow-electron scattering in solid surfaces. In addition, there is work on elastic and inelastic interactions between electromagnetic radiation and atoms and solids involving excitation of inner electrons, especially as they apply in the x-ray region and on the phase-sensitive coupling of optical fields in solids to suppress noise. Finally, the nonlinear coupling between x-rays and other waves is studied in the domain of dynamic x-ray modes.

Another, and new, aspect of solid state studies involves the theoretical and experimental investigation of distributed parameter networks on integrated electronic circuits. There has always been a search for ways to utilize an existing technology to its fullest advantage. By using a distributed parameter viewpoint on silicon planar technology, it is possible to synthesize functions on chips which are much smaller and less expensive than chips containing lumped elements only. The result of initial tests have shown the feasibility of manufacturing totally-integrated voltage-tunable RF amplifiers, IF strips, FM demodulators, oscillators, and other frequency-selective electronic networks. The elimination of large lumped inductors itself results in an appreciable saving in size, weight, material, and ultimately, cost. The design is simple and well suited to mass production techniques in silicon technology.

F. WAVE-MATTER INTERACTIONS

Program Director: N. Marcuvitz

Several different programs comprise the area of wave-matter interactions. The most comprehensive and challenging of these programs is a relatively new one which involves the interaction between very high power electromagnetic waves and materials, in which the interactions are highly nonlinear and turbulent. Other programs have been underway which treat interactions appropriate to more moderate energy levels; such interactions involve quasi-linear processes which can be treated either by equations possessing weak nonlinearity or by linearized equations, the latter leading to a large variety of parametric processes which have been studied by us in detail over a period of time. The above-mentioned studies are primarily theoretical but they have experimental phases as well. Other programs include various experimental and theoretical investigations of plasma wave properties and plasma turbulence effects, and studies in space radiophysics, which include terrestrial and extraterrestrial phenomena.

A comprehensive program has been initiated to study phenomena associated with the interaction of high power density electromagnetic waves and materials. Its motivation is to extend known techniques on linear wave propagation in spatially inhomogeneous and time varying deterministic media to practical structures wherein the material and waves are nonlinear, inhomogeneous, time varying, and turbulent. Of particular interest are self-consistent studies of reflection, transmission, and mode conversion properties of high power waves as a function of frequency, pulse width (spatial and temporal), etc., for the range of power densities wherein material properties become markedly nonlinear and experience phase changes. Preliminary model computer studies are being carried out with the aid of an interactive graphic interpretive (IGI) language developed for a PDP 11 computer facility; related experimental and theoretical investigations are also under study.

Another program, which is also relatively recent, involves weakly-nonlinear interactive systems; the approach adopted employs a scattering formulation applied to nonlinear transmission lines. A recent result, involving the general problem of reducing a nonlinear equation to a linear one, has been to construct, in a simple and straightforward manner, a general such transformation for a class of generic equations, of which

the well-known Burger's equation and the K-deV equation are special cases. Another important result has been to show on a rigorous basis that the energy in a lossless linear transmission line is conserved, in contradiction to the published literature which characterizes the presumed violation of the conservation law as a paradox.

The program on parametric processes, which involve the linearization of weakly-nonlinear interactions, has been highly successful over an extended period and now includes application to nonlinear optics and to processes involving plasmas. Recent theoretical studies have stressed the interaction of light with plasmas, including the study of stimulated Brillouin scattering (SBS) and parametric-decay instabilities. Each of these nonlinear-optical interactions involves the ion-acoustic wave in the plasma. The dependences of the instabilities on the dispersion, absorption, anisotropy and inhomogeneity of the medium are of particular concern. The results of this investigation have practical applications in laser-produced plasmas for fusion and ionospheric probing by radio waves. Other applications of parametric interactions are found in the development of new coherent optical sources for communications, and in design of pulse-echo devices for data processing. Theoretical studies have employed a Floquet expansion procedure which takes all the space-time harmonics into account, and is rigorous within the small-signal regime. The rigorous boundary-value problem and the initial value problem for instability evolution for an intense light beam incident on a medium are being solved and applied to a variety of physical cases. Instability thresholds in unbounded regions have been further investigated for interesting cases where two interactions (involving three frequencies) merge into one interaction (involving four frequencies) upon application of a small change in model parameters. Such merger of interactions appears likely in certain parametric interactions taking place in gradient inhomogeneities. The study of a class of problems dealing with intracavity parametric interactions has been initiated, prompted by interesting experimental results obtained in mode locking a ruby laser using stimulated Brillouin scattering in a birefringent crystal. In other experiments, the dynamic build-ups of SRS and SBS phonons in crystals are being studied on a time and space resolved basis in order to illustrate some of the theoretical predictions mentioned above.

The research activities related to plasmas include the examination of plasma waves and the study of plasma turbulence. Both theoretical investigations and laboratory experiments are being conducted in these areas. Both linear properties and nonlinear effects of plasma waves have been and are being examined, and related experiments are being performed to verify the theoretical findings. Among the topics which have been treated are: a) the propagation characteristics of electrostatic plasma waves in a magneto-plasma; b) plasma heating using the resonance damping of cyclotron waves; c) evolution of parametrically-excited ion-acoustic instabilities and plasma heating; d) propagation of kilovolt-subnanosecond base-band pulses along a plasma column; e) mixing of electromagnetic waves in a plasma through modulation of the electron temperature; f) coupling effects between electromagnetic modes and electrostatic modes; g) examination of the wave properties related to microwave interferometry in finite-size plasmas; and h) resonance excitation of plasma waves by a slotted cavity. The effort in this area is being continued and coordinated with the study of plasma turbulence.

The study of plasma turbulence has been gaining more emphasis recently. A comprehensive theoretical restructuring of plasma turbulence has been carried out, and new experimental facilities were designed and built to perform sophisticated experiments on plasma turbulence. One of the new experimental facilities is a long, hollow-cathode arc system, which can produce both quiescent and weakly-turbulent highly-ionized plasmas. Not only were the static characteristics of weak plasma turbulence studied, but the dynamic effects of the transition from the quiescent state to the weakly-turbulent states were also observed. Advanced data processing techniques were developed to gain reliable data on plasma fluctuations. One of the distinguished achievements in this area

is the clear quantitative evidence of the plasma turbulence generated by the electrostatic ion-cyclotron waves and of the turbulent diffusion process. Dynamic or feedback stabilization techniques are being applied to isolate the various terms contributing to the growth rate of the instability and the transition into the weak turbulence regime.

Associated with the investigation on plasma physics is the research in space radiophysics of natural terrestrial and extraterrestrial phenomena. Emphasis is on the physics of the upper atmospheres and ionospheres of the planets and their interaction with the sun's emanations. The program involves theoretical and experimental studies of disturbances in the ionosphere as produced by man-made events and by natural phenomena, such as thunderstorms. Our facilities permit the sounding of the ionosphere for disturbance detection. Radio emissions at VLF as detected by satellites have also been studied. The problem of the coupling between waves in the neutral atmosphere and waves in the ionosphere is presently being studied. Resonances have been discovered and it is planned to seek confirmation both from spacecraft data and from ground-based measurements. Dynamic interaction of the atmosphere and ionosphere of planets due to heating by solar radiation has also been studied from the standpoint of the outflow of gases and ionization as related to the formation of the solar system. These methods have been applied to the Saturnian Satellite Titan and to Jovian satellites, which are unique in the Solar System.

G. ELECTRIC POWER ENGINEERING

Program Director: E. Levi

The recent energy problems and the increased awareness of the deterioration of the environment are likely to bring about a significant increase in the share of energy utilized in electric form. In response to these mentioned needs, the Polytechnic has initiated a new research program, and has enriched its academic curriculum by instituting electric power engineering options at both the undergraduate and graduate levels.

Under investigation are many critical areas in the generation, transmission, and distribution of electrical power. Of particular interest are: (1) the magnetic separation of radioactive isotopes in nuclear fuels and waste products; (2) the application of novel technologies to the development of individual components, such as generators, high voltage transmission lines, gas insulated substations, circuit breakers, fault-current limiters, rectifiers and inverters; (3) the stability, reliability, and economy of large integrated systems; and (4) their effect on the environment and the quality of life.

At the utilization end, transportation has been singled out as the primary area of concern, since it consumes about one quarter of the overall U. S. energy budget and one half of the oil, and since it creates serious pollution problems. In the U. S. only one percent of the railroad is electrified, as compared with more than 90 percent in the other industrialized countries of the world. Besides seeking improvement in passenger and freight train services, new modes for individual, as well as mass transportation are being developed. These include electrified urban thoroughfares, as well as highways. A significant effort is devoted to linear propulsion by means of iron-cored synchronous motors. These motors present challenging problems in electromagnetics because of their complex geometries; a recent study involves a new method for the determination of magnetic fields in air gaps. Another problem which arises in such linear motors is the need to achieve a more accurate assessment of eddy current losses and field penetration in thickly-laminated iron structures. In addition, as a result of pioneering work conducted at the Polytechnic, novel topologies for flux inter-linkages have been conceived.

II. SYSTEMS

A. COMMUNICATIONS

Program Director: L. Kurz

The research program in communications and information processing currently covers topics in the optimization and evaluation of digital data transmission systems, robust detection and estimation techniques, and FM receivers.

The transmission of digital data in the presence of intersymbol interference and noise is a classical problem in communications. The problem has recently taken on an even more important significance as data rates become increasingly higher and, therefore, the effects of intersymbol interference and noise more severe. The main research effort in this area was concentrated on the optimum and suboptimum signalling and detection in the presence of intersymbol interference and gaussian noise, detection in the presence of noise represented by a mixture distribution model, and techniques useful in the monitoring of channel conditions. In particular, a new design procedure for recursive minimum mean-square error equalizers was developed, two types of detectors using digital filters were investigated, suboptimum detectors for signals corrupted by gaussian and impulsive noise were analyzed, and two classes of direct and indirect robust estimators of channel conditions were introduced and compared to existing procedures. Work is continuing on all the above-described areas of digital data system optimization.

Considerable interest generated in detection, feature extraction and estimation problems, where little is known about the data to be processed, motivated the development in recent years of several classes of rank and non-rank robust (insensitive to underlying distributions) detection and estimation procedures. Though the rank detectors tend to be more powerful, its non-rank competitors are more robust and simpler to implement. Thus, the main research effort was concentrated on developing further the latter class of detectors stressing applications in the detection of non-constant signals in mixture distributions and fading environments, sequential and nonsequential detection of underwater sounding and two-dimensional data. A parallel effort included the development of robustized estimation techniques which are applicable in adaptive modes of operation in data processing. In particular, a new family of variable-threshold non-rank procedures was introduced and compared to competitors based on the rank procedures; slippage and analysis of variance techniques were modified to operate on an adaptive mode, a new class of quadratic three-sample partition detectors, which is useful on change-of-scale and stochastic ordering detection problems, was generated and shown to be an effective and easy-to-implement class of competitors to the rank detectors; various techniques were developed for finding optimum scores and thresholds for partition detectors; two basic theorems pertaining to robustized recursive estimation were proven. Work in all the areas of robust detection and estimation described above will be continued.

Another study is concerned with investigating a novel FM demodulator capable of combatting the effects of interfering signals. This demodulator, consisting of two phase-locked loops interconnected in a feedback arrangement, has been constructed and is being studied. Further theoretical and experimental analysis will continue.

B. COMPUTERS AND COMPUTER-COMMUNICATION NETWORKS

Program Director: E. J. Smith

The research program in Computers and Computer-Communication Networks covers a number of topics directed toward the development of improved computer architecture, improved languages or information structures, techniques for implementing algorithms, data networks, and message switching systems.

In computer communications, we are concerned primarily with techniques for the better understanding of and improved design of data and computer networks. The stress is on message store-and-forward networks with minicomputers used to combine or concentrate incoming messages, to buffer or store them if need be, to route messages either dynamically, or following prescribed routing algorithms. There are interesting questions here of the appropriate modeling of interconnected networks of computers to carry out the necessary analytical work and of the appropriate choice of message statistics to be used in the modeling. The combining or concentration function is being studied, with comparisons made between different combining techniques. Dynamic buffer schemes are being studied, with optimum buffer size and comparison of various schemes a specified goal. From the overall network viewpoint a comparison of various adaptive routing algorithms is underway, as well as methods for maintaining flow control throughout the network. Also being studied are algorithms for network design taking into consideration topology, capacities, routing, reliability, and other factors.

Three specific projects have been completed which deal with adaptive routing and resource allocation, multiple routing networks, and resource-sharing in computer-communication nodes. The present effort focuses in particular upon teleprocessing network design algorithms, channel assignment in mobile telecommunication systems, and adaptive message-switching networks.

Investigations of the dynamic behavior of computer-driven communication systems are concerned with the development of models which might be useful as analytical tools in the evaluation of such systems as well as in the ultimate design of improved systems. In order to provide a focus for the effort, a particular message-switching system is studied in detail and the approach is directed toward the development of an interconnectivity model in which the system is viewed as a collection of program modules, each of which communicates with other modules through an interconnection medium via a particular machine. This often obscures the original intent of the designer of the system and provides no insights into the necessary data structures and operations for performing message-switching functions. Motivated by these considerations, current work is directed toward the development of a higher-level algorithmic description language for use in the design and specification of switched communication systems. Desirable characteristics of such a language are defined and, as an initial step, the ideas are applied to the description of a communications scanner channel. The modules are characterized by various parameters such as core size, execution time, data base, communication linkages to other modules, etc.; and the application of the model is viewed with respect to the evaluation of system changes and debugging difficulty, throughput, and local optimization of individual models. A first version of a communications-oriented design language has been completed and a current effort is concerned with the investigation of appropriate data structural forms for efficient implementation of the language. Also recently completed was an approximate analytic model of a message-switching system which provides the capability of predicting message-processing delays from a knowledge of the message switch architecture and the traffic statistics, as well as permitting one to estimate the effect on performance of transferring a task from one processor to the other in a dual-processor system. The current effort attempts to extend the model to a multiprocessor system consisting of microprocessor modules.

In the area of switching theory, a dynamic fault-test generation scheme has been developed for combinatorial logic circuits, and an extension of the same technique is being applied to the case of an asynchronous sequential machine in which tests are generated from the acyclic circuit after a near-minimum number of feedback cuts have been made in the original circuit. Dynamic tests are sought rather than static tests since the former do not require checking and consequently require less complex computation. Techniques are also being investigated for the determination of partitions having the substitution property, or other properties of special interest, through use of the predicate calculus; efficient algorations for computation are sought.

In the area of machine architecture, several studies are in progress, including the exploration of a simple, low-cost, bus-oriented computer having a common memory and instruction format for microprograms and programs as an array processor, and the investigation of a variable-precision arithmetic technique in which all numbers are stored internally within the computer as quotients of integers. The resulting effect upon computational accuracy and machine time will be studied; other related schemes will be sought out and explored.

C. SAFETY, RELIABILITY AND SOFTWARE ENGINEERING

Program Director: M. L. Shooman

The areas of Safety, Reliability and Software Engineering encompass a broad spectrum of analytical modeling, experimental systems research, and systems engineering. The underlying thread of cohesion in these areas is the application of modern techniques of probabilistic modeling, statistical measurement and experimentation, and the development of design and optimization techniques within the engineering process. The emphasis is at times on the basic development of the tools and methodology for such studies and often on the application of these new as well as existing techniques to the advancement of the state of knowledge in the given area.

This section describes and reports the work of several diverse groups within the Institute, some who are developing technology in areas outside of electronics, but with direct application to electronic devices and systems, and others who are working within the mainstream of electronics.

To the probabilistic modeler, in the abstract, the concepts of safety, reliability and availability have a unifying thread. The probability of no equipment, human, or procedural failure which endangers a human being is expressed in the safety index. Similarly, if the failure described in the preceding sentence affects system operation so as to cause a failure we are discussing reliability. Clearly, safety and reliability analysis have the same methodology; however, the definitions, implications, and design goals differ significantly. The availability index is a probability which measures the percentage of systems which are up at any specified time and allows one to model and measure the effectiveness of failure-repair dynamics. Maintainability is measured by repair rate, and repair and failure rate combine through the system configuration to determine availability. Lastly, when we turn our attention to software we find that the concepts of system reliability, availability, and maintainability apply well; however, it is important to emphasize the differences between hardware and software in constructing our models.

Activities in the safety area have recently been stimulated and accelerated by the arrival of new funding from the Department of Transportation for the second year of a proposed three-year intermodal study of transportation system safety. This work is an interdisciplinary effort with several participating departments. Clearly, although much

of the transportation vehicle and its guideway involves mechanical systems, a very high and increasing percentage of modern ships, aircraft, rail, and automobile systems involve electronic control, communications, sensing, guidance, and so on. One of the chief methodological tools being applied to this area is the system fault tree. The most important theoretical aspects are its construction, computerization, and collection of probabilistic input data. In this latter area some new techniques of statistical estimation called consensus estimation are being developed. These differ from, but bear resemblance to, the Delphi techniques used in forecasting. Also of very great importance is the modeling of the human operator. In a transportation system it is impossible to divorce the human operator from the control system. Although a great deal has been done to model the human transfer function, researchers have had less success in modeling human errors which vitally affect both reliability and safety. We have enlisted the aid of an experimental psychologist to work with the systems engineers in this area of research and hope that the cross fertilization of ideas will lead to new approaches.

The reliability (including availability and maintainability) area has been a focus of activity for many years. Many topological methods of system analysis have been developed along with approximate bounds which have served as the basic algorithms for many of the computer analysis programs produced in recent years. Also, much has been done in applying Markov models to a wide variety of problems in availability computation. Recent thrusts involve (1) a queueing theory approach to availability and maintainability which allows more complex repair models; (2) modeling of the inspection interval problem (applied to critical systems), so as to optimize the benefits in safety and minimize the cost; and (3) the study of computer-assisted test of electronic systems with regard to design constraints, optimum test points, and overall cost minimization and reliability maximization.

In the area of software engineering, a substantial portion of this effort is supported by the Rome Air Development Center. Activities in this area are focused upon five aspects of software reliability: a) construction of probabilistic models for software errors, which reflect the content and type of error, removal and generation rate, prediction of mean time between failure and reliability; b) measures of computational complexity based upon the algorithm, automata theoretic or graph theoretic complexity, information content and program size; c) relationship of modular and structured programming styles to reliability; d) modeling techniques for program verification and testing including optimal test strategies for nonexhaustive tests; e) models for the comparison of programming languages, comparative study of formal definition languages, and techniques for compiler writing.

The probabilistic modeling effort dealt with the construction of many-state Markov models for the determination of software availability and macro models to predict number of bugs and program reliability. Preliminary models were developed to predict operational software reliability. In the second phase of this work, error-generation terms were added to increase the degree of realism. A new micro approach relates errors and reliability to a functional path decomposition of the program.

Several approaches are being taken to provide measures of complexity, both for the problem and the ensuing software. The initial efforts treated the complexity of a function via recursive function theory. Present efforts are directed toward relating program complexity to established ideas in the fields of automata, communications and information theory. Recent results provide a measure of program length based on Zipf's Law and operator/operand count.

In the area of structured and modular programming two efforts are being pursued. One effort, presently in progress, applies modular programming techniques to the automatic-interactive construction of programs. In addition, the statistical design of an experiment for the objective evaluation of structured vs. nonstructured programs is in the planning phase.

In the program test area, two efforts are in progress. The objective of the first effort is to develop a method which computes the number of tests necessary to verify a computer program. Verification is categorized into several classes, including "exhaustive" at one extreme and "processing of each instruction at least once" at the other. In the second effort, an analytic method is being developed for the selection of data for automatic program testing.

In the area of languages, a new effort has begun to construct a very high level language for writing program specifications. It is envisioned that this language will lie between a high level programming language (FORTRAN, PL/1, etc.) and English prose, but will be concise and yet avoid ambiguity.

D. SYSTEMS, CONTROL AND NETWORKS

Program Director: D. C. Youla

Research activity in the network area encompasses both the classical and modern lumped-distributed domains. In the latter, several significant breakthroughs have been made which are expected to lead to an exact insertion-loss synthesis technique for the design of optimum filters incorporating cascades of equi-delay TEM lines and lumped, lossless two-ports. In particular, it is expected that transformers and filters exhibiting equi-ripple performance in both pass and stop bands can be designed to specifications. In principle, there appears to be no reason why the method cannot be extended to single-moded waveguide networks employing obstacles to produce the lumped discontinuities. An unexpected and important outgrowth of the above study has been the development of a new diagnostic digital technique for probing one-dimensional nonuniform structures. Hopefully, the modeling of dispersion by means of lumped two-ports will enlarge the scope of the method significantly.

A major topic in the control area involves the application of classical ideas to the design of feedback controllers for linear multivariable systems. All efforts on this topic have been completely successful. A totally new frequency-domain technique has been developed for the design of optimal multivariable controllers. The class of problems that fall within the scope of this technique is much broader than that encompassed by the linear, quadratic gaussian approach (LQG). The latter revolves around the idea of Kalman filtering and for this reason is unable to absorb in a natural and straightforward manner many essential practical constraints. The group is presently engaged in work which should result in an effective computer implementation of the optimal controller. The availability of such an algorithm will undoubtedly suggest related simpler suboptimal strategies and open the door to significant industrial applications.

Work on the control of stochastic systems has been focused on situations in which discrete events, such as equipment failures or message arrivals, occur at random times. Recursive optimization equations have been derived in many cases via dynamic programming. Novel techniques have been devised to solve those equations to get optimal decision rules for diverting messages or vehicles to alternate routes in situations with simple arrival and service time probability distributions. In more complex cases, structural properties of the optimal controllers have been derived and specific control algorithms have been compared via simulation.

In the area of reliability applications, both repairman assignment and optimal inspection problems have been solved. The problems studied here have been of a higher order of complexity than those generally formulated by system analysts. In one case, both feedback control for dynamic tracking accuracy and repairman assignment policies were simultaneously optimized for repairable systems. In another, techniques were developed for optimizing inspection schedules when the corresponding tests degrade the lifetimes of the very components whose status is being checked.

Identification and parameter estimation from random data is of central interest in systems applications. In control applications, one must have better knowledge about the plants as well as effective models of certain plant elements or functions for which there are no intrinsic analytical descriptions available. This is especially true when human elements are in the loop and we try to describe neuromuscular systems such as arm and leg control, pupil response, etc. We also require identification and modeling techniques when the system itself is too complex so that its fundamental input-output properties are unknown. This is especially true of large-scale economic systems where dynamic models are obtained in some optimal fashion from available data in order to describe the evolution of the system as a result of changing policies. Finally, even in the case for which dynamic equations describing a system are known or accepted, parameters are often unknown and must be estimated. Identification and estimation techniques have been applied to all of the problems described above. But, to a great extent, the basic feature of previous applications of identification and parameter estimation is that the models of the systems studied have been time invariant, and their parameters have been estimated from statistically stationary data (after trends have been removed). However, there exist important phenomena for which, even after trends have been removed, the data is intrinsically non-stationary, characterized usually by a period of increasing and then decreasing intensities. Such characteristics are found in economic systems, but are even more prevalent in geophysical observations such as meteorological fluctuations and geological phenomena such as earthquake excitations. Motivated by this last application to characterize the statistical properties of earthquake excitations, recent studies have led to new limit theorems for estimating significant parameters in models of non-stationary time series, and has led to a specific approach for modeling earthquake acceleration data. This approach has been implemented on computers and is being applied to study recent earthquake acceleration data obtained from the Western United States. We are just at the beginning in developing techniques for the identification of non-stationary systems and data. Many problems remain to be solved.

E. DATA PROCESSING

Program Director: A. Papoulis

Current studies in data processing include the following aspects of picture (or tabular) processing and spectral estimation: statistical enhancement and extraction techniques for pictorial or tabular data; reduction of diffraction effects in the imaging of coherent and incoherent objects; image distortion resulting from atmospheric turbulence; and statistical analysis in spectral estimation.

Statistical enhancement and extraction techniques for pictorial or tabular data: This research concentrates on developing procedures to present the data in a useful form or to present the data in some way in which factors in the data which are "almost invisible" become plainly visible. The effort was concentrated on factor analysis and masking techniques. The masking operation was taken to mean the process in which a "window" or "mask" is swept across the data either electronically, mathematically,

mechanically, or some combination of these. The motives for selection of a particular mask are simplicity and effectiveness. The preliminary results indicate that both approaches to the enhancement and extraction problems of tabular data are promising, and the research effort utilizing these and related techniques will be continued.

Reduction of diffraction effects in the imaging of coherent and incoherent objects: The known methods of deconvolution in one and two dimensions for analog and digital data are compared for accuracy and computational economy. A new technique is developed for the complete recovery of objects of finite size. The technique is based on an iteration scheme involving only the FFT. It is shown that, in the absence of noise, the iteration converges to the unknown object. The effects of noise and the aliasing errors are determined and it is shown that they can be controlled by early termination of the iteration.

Image distortion resulting from atmospheric turbulence: In the recent literature, a method has been proposed for a dynamic correction of this distortion. The underlying filtering (Poisson noise) and prediction (control delay) problem leads to a two-dimensional Wiener-Hopf equation whose solution is under investigation under various assumptions about the spectrum of the turbulence. The reverse problem of determining the properties of the medium in terms of image of a moving sattelite has also been considered.

Statistical analysis in spectral estimation: The current emphasis is on the statistical analysis of the method of maximum entropy and in comparison with other methods for various special forms of the unknown spectra. An adaptive perturbation scheme is under consideration for updating the estimated spectra as new information is received. The convergence of the scheme to the smoothed spectrum is under examination.

POLYTECHNIC INSTITUTE OF NEW YORK

G. Bugliarello, President
J.J. Conti, Provost
R.J. Cresci, Associate Provost for Research

B.J. Bulkin, Dean of Arts and Sciences
L.B. Felsen, Dean of Engineering
A.B. Giordano, Dean of Graduate Studies

MICROWAVE RESEARCH INSTITUTE DIRECTORATE

A.A. Oliner, Director

J. Fox
S.W. Rosenthal

FACULTY AND RESEARCH STAFF PARTICIPATING IN MICROWAVE RESEARCH INSTITUTE PROGRAMS

DEPARTMENT OF CHEMISTRY

Faculty

E. Pearce: Department Head

E. Banks
J.M. Casper
P. Hoggard
N.C. Peterson

Postdoctoral Fellows

R. Carvalho
M. Greenblatt
C. Kosky
S. Nakajima

Fellows, Trainees and Research Associates

M. Shone
G. Torre

DEPARTMENT OF ELECTRICAL ENGINEERING AND ELECTROPHYSICS

Faculty

T. Tamir: Department Head

M. Adamowicz
L.B. Bergstein
H.L. Bertoni
J.J. Bongiorno, Jr.
R.R. Boorstyn
F.A. Cassara, Jr.
E.S. Cassedy
B.R. Cheo
D. Davids
R. Drenick
H. Farber
L.B. Felsen
A.B. Giordano
S.H. Gross
A. Hessel
F. Kozin
L. Kurz
A.E. Laemmel
J.T. LaTourrette
E. Levi
N. Marcuvitz
E. Mishkin
M.C. Newstein
A.A. Oliner
I. Palócz
A. Papoulis
S.W. Rosenthal
H. Ruston
H. Schachter
B. Senitzky
L. Shaw
J. Shmoys
M.L. Shooman
L.M. Silber
E.J. Smith
S.G. Tyan
W-C. Wang
D.C. Youla

DEPARTMENT OF ELECTRICAL ENGINEERING AND ELECTROPHYSICS

Research Staff

B. Rudner, S.Y. Shin, N. Solimene, W.T. Walter

Research Professors

L. Birenbaum, E-W. Hu, T. Parsons, S.T. Peng, Z. Zabar

Postdoctoral Fellows

S. Choudhary, H.J. Eun, M. Jain, N. Kazanjian, T.S. Lee

Fellows, Graduate Assistants and Graduate Students

H. Ahn
H.S. Ashtiani
A. Benler
S. Brutzkus
P.K. Bondyopadhyay
A. Camello
C. Chamzas
C.Y. Chang
K.C. Chang
S. Chikte
Y. Devedas
A. D'Sa
A.R. El Fiqi
J. Goldfinger
K.I. Haddad
S. Has
C. Hechtman
A. Heuman
C.L. Hsu
H.M. Huang
T. Ishihara
M. Kavehrad
N. Kazanjian
R.T. Kinasewitz
P.P. Krzykos
H.M. Kudyan
G. Kull
S-P. Kuo
F. Lalezari
A.J. Lieberman
E. Lipshitz
E. Mayer
J.H. Mirza
F. Nakajima
V.K. Nangia
S. Natarajan
S-T. Ng
R. Parekh
S-K. Park
R. Pascone
M. Petrella
G. Ramanathan
J. Rosenberg
L. Rosenheck
J. Sagiv
Y. Saito
A. Sanchez-Corpas
V.S. Shah
S. Siddiqi
W. Simmons
A.M. Simonoff
D. Singh
S-K. Sohraby
J. Sokol
C.C. Su
B. Tomasic
S.G. Tsapepas
D.H. Vapheas
E. Voudouri
K-C. Whang
M.H. Wu

DEPARTMENT OF PHYSICS

Faculty

T. Kjeldaas: Department Head

H.J. Juretschke, W. Kiszenick, B. Post, D.B. Scarl, H.W. Schleuning

Research Scientist

S.R. Barone

Research Associate

M. Eschwei

Fellows, Trainees and Graduate Assistants

S.L. Chang
R. DeSerio
L. Goldstein
H. Grossman
J. Gutreich
T. Hom
D.J. Lischner
P. Mazumdar
G. Osche
P. Persans
J. Selby
S.M. Turner
P. Wang

MICROWAVE RESEARCH INSTITUTE

MRI Proceedings

Editorial: J. Fox, Editor

MRI Technical Reports

Annual Report Coordination: J. Fox J. Maher

Editorial: J. Maher S. Mahrous

Graphic Arts: R. Gnaffo

WORD PROCESSING CENTER

M. Fischetti, Supervisor

L. Babikian C. Devlin R. Drucker
E. Cummings E. MacDonald

I. ELECTROPHYSICS

A. ELECTROMAGNETICS

Program Director: A. Hessel

B. ACOUSTICS

Program Director: H. L. Bertoni

C. OPTICS

Program Director: L. Bergstein

D. QUANTUM ELECTRONICS

Program Director: W. T. Walter

E. SOLID STATE AND MATERIALS

Program Director: H. J. Juretschke

F. WAVE-MATTER INTERACTIONS

Program Director: N. Marcuvitz

G. ELECTRIC POWER ENGINEERING

Program Director: E. Levi

MUTUAL COUPLING IN ARRAYS ON CONCAVE SURFACES

H. Ahn and A. Hessel

This work is devoted to the development of the ray analysis of coupling coefficients in collector arrays of aperture elements on the interior concave surfaces in scannable feed-through lenses of the Dome Antenna[1] type. In such a lens, the energy is radiated by a planar feed array which is located on the diameter of a passive lens shaped as a dome. This energy is received by the elements of the collector array located on the interior concave side of the dome and is transmitted via a bank of passive phase shifters to the radiator elements placed on the exterior convex side of the dome. To assure an efficient energy transfer, both the collector and the radiator elements must be matched in their respective array environments for an appropriate range of incidence angles and over a desired frequency band. To ascertain the degree of mismatch the designer usually avails himself of experimentally determined coupling coefficients, which, in turn, for a given excitation determine the collector and the radiated array efficiency (mismatch loss). The virtues of the ray method which is being developed for arrays on concave surfaces are that, in contrast with the modal techniques, a) they are not restricted to a separable array surface geometry and b) they permit the exploitation of the planar phased array computer programs. The ray fields employed here are not of the usual type in the presence of a perfectly conducting surface, but rather represent locally plane waves radiated by the excited element and received by a collector element, weighted by their patterns in a locally planar, matched array environment. These locally plane waves are received either via a direct ray or after one or more internal reflections. The ray internal reflection coefficients correspond to a plane wave impinging, in the ray direction, on a locally planar array with the geometry appropriate to the point of reflection.

The collector array model is a uniformly spaced circular array of N open ended "single mode" parallel-plate-guides in a concave cylindrical surface (Figure 1). The waveguides propagate only the dominant mode of the appropriate polarization, and are equipped with a network that would match a corresponding planar array at broadside. Modal analysis in a unit cell yields the active (phase sequence) aperture admittance $Y_a(n)$ and reflection coefficients $\Gamma(n)$ in the form

$$\Gamma(n) = \frac{Y_p(0) - Y_a(n)}{Y_p^*(0) + Y_a(n)}, \quad (n = 0, \pm 1 \ldots \pm N/2) \; . \tag{1}$$

For TM case,

$$Y_a(n) = j\sqrt{\frac{\epsilon_o}{\mu_o}}\,\frac{b}{d}\sum_{m=-\infty}^{\infty}\left(\frac{\sin(n_m b/2r_o)^2}{n_m b/2r_o}\right)\cdot J_{n_m}(kr_o)/J'_{n_m}(kr_o) \tag{2}$$

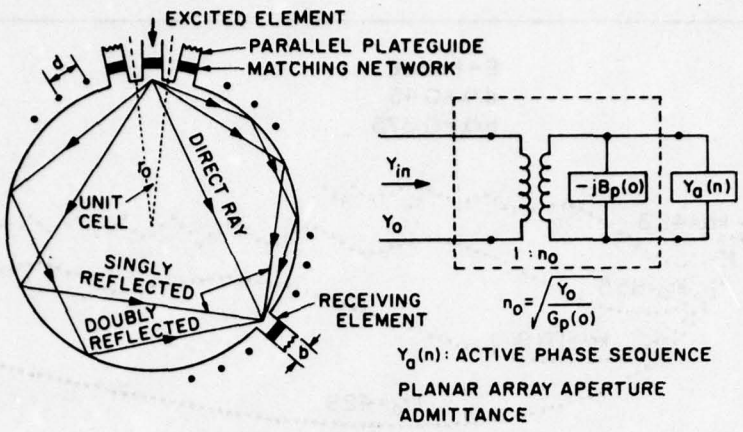


Fig. 1. Cylindrical collector array model, ray structure and matching network.

and for TE case,

$$Y_a(n) = -j\sqrt{\frac{\varepsilon_o}{\mu_o}}\left(\frac{2b}{d}\right)\sum_{m=-\infty}^{\infty}\frac{\cos^2 n_m b/2r_o}{\left(\frac{\pi}{2}\right)^2 - \left(\frac{n_m b}{2r_o}\right)^2}\,\frac{J'_{n_m}(kr_o)}{J_{n_m}(kr_o)} \quad . \tag{3}$$

The coupling coefficients are obtained from Eq. (1) via

$$S_\ell = \frac{1}{N}\sum_{n=-N/2}^{N/2}\Gamma(n)\, e^{-j\frac{2\pi}{N}n\ell} \quad , \quad \left(-\frac{N}{2} \le \ell \le \frac{N}{2}\right) \quad . \tag{4}$$

Here, $n_m = n + mN$, $Y_p(0) = G_p(0) + jB_p(0)$ and $J_n(kr_o)$ is the Bessel function of the first kind and nth order. Samples of numerical results based on Eqs. (1) - (4) are shown in Figure 2. They indicate that for $d/\lambda < 1/2$ the $|S_\ell|$ curve follows initially that of a planar array, but that gradually the decay rate of $|S_\ell|$ decreases, the curve $|S_\ell|$ vs. ℓ passes through a dip, with a subsequent rise, accompanied by a distinct ripple. These features are similar for both the E- and H-plane coupling coefficients.

To develop the ray method, Eqs. (1) - (4) for TM are cast into a form that permits an asymptotic evaluation of S_ℓ. This is accomplished by imbedding the cylindrical collector array in an infinite angular space.[2,3] As a result, the unit cell Floquet phasing ν is no longer limited to integral values as in Eqs. (1) - (4), but becomes continuous. Equations (1) - (4) transform into

$$\Gamma(\nu) = \Gamma(|\nu|) = \frac{Y_p(0) - Y_a(\nu)}{Y_p^*(0) + Y_a(\nu)} \quad , \tag{5}$$

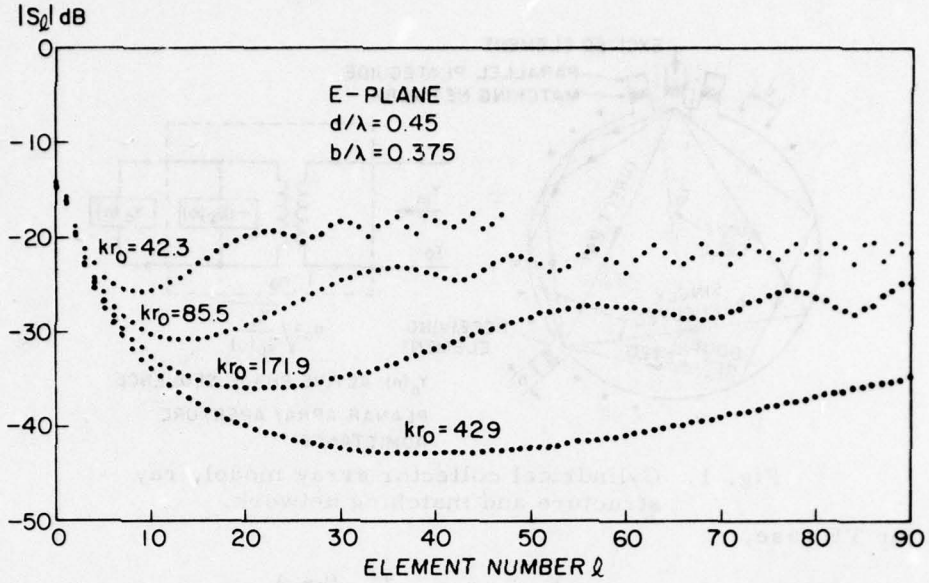


Fig. 2. E-plane coupling coefficients vs. collector array radius.

$$Y_a(\nu) = j\sqrt{\frac{\varepsilon_o}{\mu_o}} \sum_{m=-\infty}^{\infty} \left(\frac{\sin \nu_m b/2r_o}{\nu_m b/2r_o}\right) \frac{J_{|\nu+mN|}(kr_o)}{J'_{|\nu+mN|}(kr_o)} \tag{6}$$

and

$$S_\ell = \frac{1}{N} \sum_{p=-\infty}^{\infty} \int_{-N/2}^{N/2} \Gamma(\nu)\, e^{-j\nu \frac{2\pi}{N}(\ell + pN)}\, d\nu \tag{7}$$

In Eq. (7) the index p corresponds to the pth image source in the infinite angular space. Using the Debye approximation for $J_{|\nu+mN|}$, the integrand in Eq. (7) is expanded in a Neumann (multiple scattering) series (actually only a finite number of terms is permitted). After a stationary phase evaluation, each term yields a ray contribution to S_ℓ. The n times reflected ray contribution from the p = 0 source is

$$S_\ell^{(n)} = -\frac{\sqrt{2\pi d}}{\lambda} \frac{[T_p(k \sin \theta_\ell^n)]^2 [\hat{\Gamma}(k \sin \theta_\ell^n)]^n}{} e^{-jkr_o 2(n+1)\cos \theta_\ell^n} \cdot e^{j(\frac{n\pi}{2} + \frac{\pi}{4})} \tag{8}$$

Here T_p is the planar, equivalent array transmission coefficient for a scan angle θ_ℓ^n off broadside. Γ is the equivalent planar array voltage reflection coefficient for plane wave incident at θ_ℓ^n, the angle θ_ℓ^n being that between the ray direction and the local array normal at the excited element. It is seen that, while the direct ray (n = 0) does not cross a caustic, the reflected rays do, as indicated by the $j^{n\pi/2}$ factor. Equation (8) can be cast into an invariant form involving the locally planar element patterns, reflection coefficients and the distance between the excited and the receiving element. Figure 3

shows a comparison between the numerical results for $|S_\ell|$ obtained by modal[4] and ray analysis. The latter includes the direct, singly and doubly reflected rays. The agreement is excellent, (in phase as well) up to the dip region, where the Airy approximation of J_ν must be employed. This is presently under study.

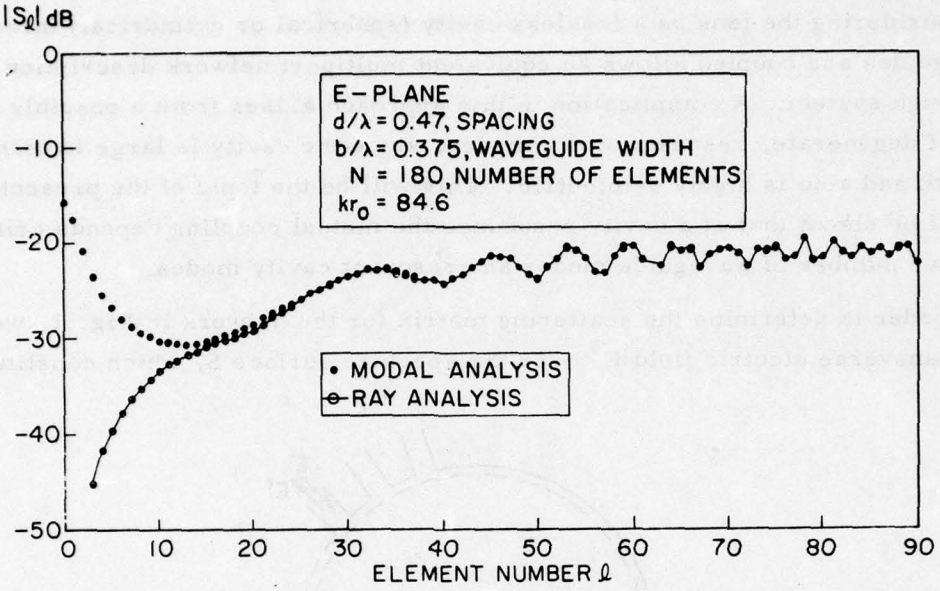


Fig. 3. Coupling coefficients in a cylindrical collector array - ray vs. modal analysis.

The ray method clarifies the formation of the $|S_\ell|$ curves. The portion past the dip is due primarily to the direct ray contribution modulated by the interference between the direct and singly reflected ray. The portion near the source corresponds to a nearly planar behaviour and the dip represents a transition region between the quasi planar regime near the source and the ray region.

Ballistic Missile Defense System Command
DASG60-76-C-0006

H. Ahn and A. Hessel

REFERENCES

1. L. Schwartzman and J. Stangel, The Dome Antenna, Microwave Journal (October 1975).
2. J. C. Sureau and A. Hessel, "Element Pattern for Circular Arrays of Waveguide-Fed Axial Slits on Large Conducting Cylinders," IEEE Trans. Ant. Prop., Vol. 19, pp. 64-74 (January 1971).
3. J. Shapira, L. B. Felsen and A. Hessel, "Ray Analysis of Conformal Antenna Arrays," IEEE Trans. Ant. Prop, Vol. 22, pp. 49-83 (January 1974).
4. A. Hessel and J. Shapira, "Mutual Coupling Coefficients in Collector Arrays on Circular Cylindrical Concave Surfaces," IEEE Trans. Ant. Prop., Vol. AP-25, pp. 741-747 (November 1977).

CAVITY RESONANCES AND MUTUAL COUPLING IN CONCAVE WAVEGUIDE ARRAYS

H. Steyskal

A systematic design of feed-through lens or dome antennas[1] requires knowledge of the mutual coupling between waveguide elements on the concave interior surface of a lens. Considering the lens as a lossless cavity (spherical or cylindrical) through which the waveguides are coupled allows an equivalent multiport network description of the feed-through system. A complication in this approach arises from a possibly large number of degenerate, resonant cavity modes, since the cavity is large in terms of wavelength and also is highly symmetric. This will be the topic of the present study and it will be shown that at a cavity resonance the mutual coupling depends critically on the relative number of waveguide modes and resonant cavity modes.

In order to determine the scattering matrix for the network in Fig. 1, we solve for the transverse electric field E^s over the aperture surface S, which constitutes the

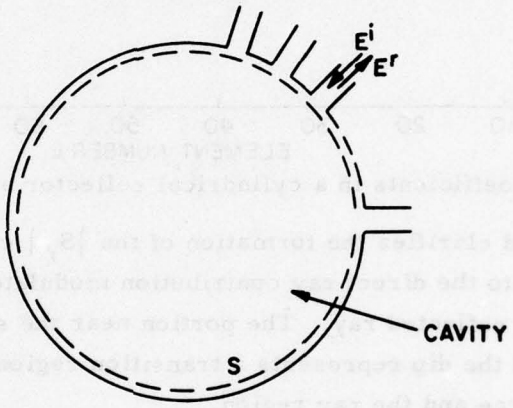


Fig. 1. Concave waveguide array -- A cavity coupled multiport network.

interface between the waveguide apertures and the cavity, in terms of the incident field E^i at S. This then gives rise to a reflected field $E^r = E^s - E^i$.

Following Galerkin's method we expand the transverse waveguide fields into a set of orthonormal waveguide modes e_n (n denoting here both the mode and the waveguide) and the cavity field into radial waves with transverse orthonormal real mode functions e_n. The aperture field E^s is approximated by a finite set of modes with unknown amplitudes v_n,

$$E^s = \sum_1^N v_n e_n \qquad (1)$$

Enforcing the continuity of the tangential electric and magnetic fields across S and the vanishing of E^s over the conducting parts of S (see Ref. 2), leads to an equation for E^s. Finally, the requirement that both sides of this equation have equal projections on the set spanned by the basis $\{e_n\}_1^N$ leads to the matrix equation

$$2 Y \bar{v}^i = C \bar{v} \tag{2}$$

Here the column vectors $\bar{v}^i = \{v_n^i\}_1^N$ and $\bar{v} = \{v_n\}_1^N$ denote the incident and the total modal voltages at S. The matrices Y and C are

$$Y = [y_m \delta_{mn}] \tag{3}$$

$$C = [y_m \delta_{mn} + \sum_{k=1}^{K} \hat{y}_k (\hat{e}_k, e_m)(\hat{e}_k, e_n)] \tag{4}$$

where y_m denotes the waveguide modal admittance, $\hat{y}_k$ the modal input admittance looking into the cavity from the aperture S and the inner product is defined by

$$(\hat{e}_m, e_n) = \int_S \hat{e}_m \cdot e_n \, dS \tag{5}$$

The solution of Eq. (2) is usually obtained by inverting the matrix C. However, when a resonance of the cavity is approached all elements of C become infinite (since some $\hat{y}_i \to \infty$) and a more careful analysis is required.

Consider a cavity with R degenerate modes so that $\hat{y}_1 = \hat{y}_2 = \cdots = \hat{y}_R$. At resonance, when $\hat{y}_R \to \infty$, we factor out $\hat{y}_R$ from C to obtain

$$C = \hat{y}_R (A + E) \tag{6}$$

where the elements of A and E are

$$a_{mn} = \sum_{k=1}^{R} (\hat{e}_k, e_m)(\hat{e}_k, e_n) \tag{7}$$

$$\epsilon_{mn} = \frac{1}{\hat{y}_R} [y_m \delta_{mn} + \sum_{k=R+1}^{K} \hat{y}_k (\hat{e}_k, e_m)(\hat{e}_k, e_m)] \tag{8}$$

and thus Eq. (2) becomes

$$2 Y \bar{v}^i = \hat{y}_R (A + E) \bar{v} \tag{9}$$

Now, the desired mutual coupling coefficient S_{pq} between the ports p and q may be determined by Cramer's rule as

$$S_{pq} = \frac{v_p^i}{v_q^i}\bigg|_{v_n^i=0,\, n\neq q} = \frac{\begin{vmatrix} a_{11}+\varepsilon_{11}, \cdots, a_{1,p-1}+\varepsilon_{1,p-1}, 0, a_{1,p+1}+\varepsilon_{1,p+1}, \cdots, a_{1N}+\varepsilon_{1N} \\ \vdots \\ \cdots\cdots 2y_q/\hat{y}_R \cdots\cdots \\ \vdots \\ a_{N1}+\varepsilon_{N1}, \cdots, a_{N,p-1}+\varepsilon_{N,p-1}, 0, a_{N,p+1}+\varepsilon_{N,p+1}, \cdots, a_{NN}+\varepsilon_{NN} \end{vmatrix}}{\begin{vmatrix} a_{11}+\varepsilon_{11}, \cdots\cdots\cdots, a_{1N}+\varepsilon_{1N} \\ \cdots\cdots\cdots\cdots \\ a_{N1}+\varepsilon_{N1}, \cdots\cdots\cdots, a_{NN}+\varepsilon_{NN} \end{vmatrix}} \tag{10}$$

In order to evaluate Eq. (10), which usually is indeterminate, when $\hat{y}_R \to \infty$, the two determinants are expanded in powers series of ε_{mn} which leads to

$$s_{pq} = \frac{\dfrac{2y_q}{\hat{y}_R}\,\bar{A}_q^p + \sum_{\substack{i,j=1 \\ i\neq p,\, j\neq q}}^{N} \varepsilon_{ij}\,\bar{A}_{qj}^{pi} + \sum_{\substack{i,j=1 \\ i\neq p,\, j\neq q}}^{N} \varepsilon_{ij} \sum_{\substack{k;\,\ell=1 \\ k\neq p,\, i;\, \ell\neq q,\, j}}^{N} \varepsilon_{k\ell}\,\bar{A}_{qj\ell}^{pik} + \cdots}{|A| + \sum_{i,j=1}^{N} \varepsilon_{ij}\,\bar{A}_j^i + \sum_{i,j=1}^{N} \varepsilon_{ij} \sum_{\substack{k,\ell=1 \\ k\neq i,\, \ell=j}}^{N} \varepsilon_{k\ell}\,\bar{A}_{j\ell}^{ik} + \cdots} \tag{11}$$

Here $|A|$ denotes the determinant of A and the cofactors $\bar{A}_{j,\ell,\cdots,t}^{i,k,\cdots,s}$ are the determinants of the matrix A with rows $i,k,\cdots,s$ and columns $j,\ell,\cdots,t$ deleted and taken with their proper signs.[3] Note that $|A|$ and all $\bar{A}$'s are constant, independent of $\hat{y}_R$.

Equation (11) is a general expression valid at all frequencies. Its particular advantage is that it readily yields the coupling coefficients at a resonance. When rank A = N the numerator vanishes as $\hat{y}_R \to \infty$, whereas the denominator is equal to $|A| \neq 0$ and consequently S_{pq} also vanishes. When rank A < N, the numerator and denominator contain the same powers of $\hat{y}_R$ since ε_{ij} is proportional to $1/\hat{y}_R$ and therefore S_{pq} will tend to a finite, non-zero value given by Equation (11).

For more insight into the resonant situation we return to Eq. (9) and note that in view of Eq. (8), $A\bar{v} \to 0$ when $\hat{y}_R \to \infty$. Writing A as the product of an R x N matrix B with elements $b_{mn} = (\hat{e}_m, e_n)$ and its transpose B^t we have

$$A\bar{v} = B^t B\bar{v} = 0 \tag{12}$$

where it can be shown, by premultiplying with $\bar{v}^t$, that the matrices A and B have equal rank. Equation (12) therefore is tantamount to

$$B\bar{v} \equiv \begin{pmatrix} (\hat{e}_1, e_1), \cdots, (\hat{e}_1, e_N) \\ \cdots\cdots\cdots\cdots \\ (\hat{e}_R, e_1), \cdots, (\hat{e}_R, e_N) \end{pmatrix} \begin{pmatrix} v_1 \\ \cdot\cdot \\ v_N \end{pmatrix} = 0 \tag{13}$$

or, in other words,

$$(E^s, \hat{e}_r) = 0 \qquad r = 1, 2, \cdots, R \quad . \tag{14}$$

As a result we thus obtain the general condition that at a resonance the aperture field E^s must be orthogonal to the set of degenerate resonant cavity modes $\{\hat{e}_r\}_1^R$. Physically this means that when the modal input admittance $\hat{y}_R$ becomes infinitely large the aperture field must not couple to the corresponding cavity mode, since it would excite an infinite H-field.

The solution of Eq. (13) depends on the rank of B. The alternatives are:

(1) rank B = N, in which case $\bar{v} = 0$ and all mutual coupling coefficients vanish

(2) rank B < N, in which case solutions $\bar{v} \neq 0$ and non-vanishing coupling coefficients given by Eq. (11) are obtained. Clearly this condition occurs whenever R < N.

However, it can also occur when $R \geq N$, namely when the set of aperture modes does not couple to one (or more) of the resonant cavity modes or some linear combination thereof. This effectively eliminates the mode from the analysis and makes the columns of B linearly dependent. Since an infinitesimal change in relative aperture positions would make them independent the corresponding solution is discontinuous with geometry. The number of such cases is limited and forms a set of zero measure (see the example below).

A conclusion to be drawn is that for a method of moment analysis of the cavity a sufficient number of aperture modes should be used in order to resolve the cavity resonance. For a circular cylindric cavity this leads to N > 2, and for a spherical cavity N > 2ka + 1, which can be a very large number, ka being the great circle measured in wavelengths.

A. An Example

Consider a two-dimensional circular cavity with two identical narrow slits as shown in Figure 2. We approximate the aperture field of each slit by one single ϕ-directed mode so that $E^s = v_1 e_1 + v_2 e_2$. Since the slits are narrow we set $e_1 = \alpha\delta(\phi - \phi_1)$ and $e_2 = \alpha\delta(\phi - \phi_2)$, where α is the appropriate normalization constant. Similarly, the transverse mode functions $\hat{e}_n$ of the cavity are purely ϕ-directed and due to the rotational symmetry they come in pairs, distinguished by a cos $n\phi$- and sin $n\phi$-variation only. The set $\{\hat{e}_n\}$ is thus composed of $\{\beta_n \cos n\phi\}_{n=0}^{\infty}$ and $\{\beta_n \sin n\phi\}_{n=0}^{\infty}$, where β_n again are normalization constants.

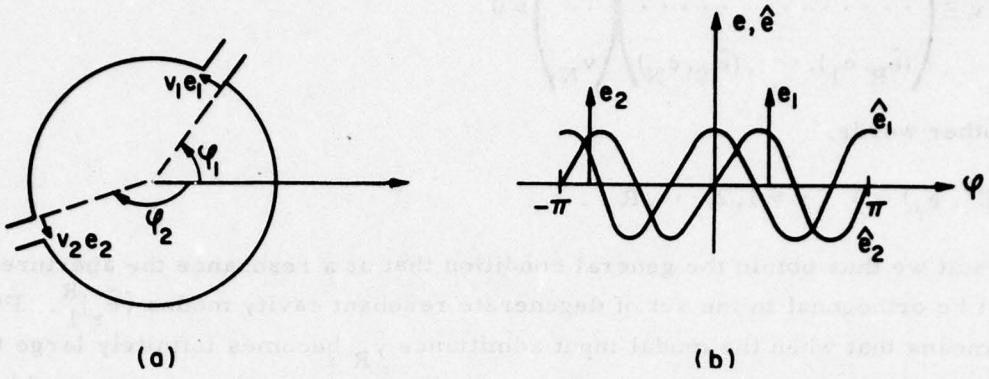


Fig. 2. (a) Two-dimensional cavity with two narrow slits, (b) Aperture modes e_1, e_2 and cavity modes $\hat{e}_1, \hat{e}_2$.

Now consider the case of modes $\hat{e}_1 = \beta_M \cos M\phi$ and $\hat{e}_2 = \beta_M \sin M\phi$ at resonance. Equation (13) becomes

$$B\bar{v} = \begin{bmatrix} (\hat{e}_1, e_1) & (\hat{e}_1, e_2) \\ (\hat{e}_2, e_1) & (\hat{e}_2, e_2) \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \alpha\beta_M \begin{bmatrix} \cos M\phi_1 & \cos M\phi_2 \\ \sin M\phi_1 & \sin M\phi_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 0$$

We see that in most cases rank B = 2, which leads to $\bar{v} = 0$ and $s_{11} = s_{22} = -1$, $s_{12} = s_{21} = 0$. However, when the slits are spaced a multiple of π/M apart the rank is reduced so that rank B = 1. Then we use Eq. (13) to obtain the scattering coefficients

$$s_{11} = s_{22} = y_1 \left\{ y_1 + \alpha^2 \sum_{\substack{n=0 \\ n \neq M}}^{K} \hat{y}_n \beta_n^2 [1 - \cos M(\phi_1 - \phi_2) \cos n(\phi_1 - \phi_2)] \right\}^{-1}$$

$$s_{12} = s_{21} = -s_{11} \cos M(\phi_1 - \phi_2).$$

U.S. Army BMDSC
DASG60-76-C-0006

H. Steyskal

REFERENCES

1. L. Schwartzman and J. Stangel, "The Dome Antenna," Microwave Journal (October 1975).

2. N. Amitay, V. Galindo, C.P. Wu, "Theory and Analysis of Phased Array Antennas," (New York: Wiley-Interscience, 1972) Chapt. 3.

3. G. Shilov, "An Introduction to the Theory of Linear Spaces," (New Jersey: Prentice Hall, 1961) p. 20.

SHADOWING OF AN INHOMOGENEOUS PLANE WAVE BY AN EDGE

H. L. Bertoni, A. C. Green and L. B. Felsen

At high frequencies, electromagnetic fields may be classified as propagating or evanescent, depending on whether their direction of propagation is real or complex. Examples of evanescent fields are those occurring on the edges of a Gaussian beam and the inhomogeneous plane wave produced in the rarer medium, as a result of total internal reflection. When either type of field encounters a reflecting obstacle of finite extent, such as a knife edge, a shadow zone is formed behind the obstacle. The formation of the shadow zone is well understood for the case of propagating fields, and can be described in terms of ray optics.

For the case of evanescent waves, the mechanism of shadow formation is not understood. In addition, the meaning of the shadow region is in question, since the incident evanescent field may be weaker in the shadow than the diffracted field generated by the obstacle. In order to study the mechanism of shadow formation and its significance for evanescent fields, we have considered the problem of inhomogeneous plane diffraction by a perfectly conducting semi-infinite screen. By a careful asymptotic treatment and error analysis of the known exact solution, the location of the shadow boundary is determined, as is the transition region surrounding it wherein the field cannot be separated into incident and diffracted constituents. It is found that for strongly evanescent fields, the shadow boundary and transition zones differ markedly from those for a homogeneous plane wave.

A. Limitation of Ray Optics for Evanescent Fields

The fields of propagating waves can be described in terms of a family of rays whose paths in space are determined by the laws of geometrical optics.[1,2] As illustrated in Fig. 1, for a plane wave, when the ray family encounters a reflecting obstacle of finite extent, such as the knife edge, some of the rays miss the obstacle and are unaffected by its presence. The portion of the ray family illuminating the obstacle is reflected. The rays separating the two portions of the incident ray family will be referred to as critical rays. Extension of the critical rays past the obstacle defines the boundary of its shadow. Treating the critical rays as if they were reflected defines the shadow boundary of the family of reflected rays (see Figure 1). In addition, the critical ray excites a family of diffracted rays, propagating away from the edge in all directions.

Except in transition regions that are of finite width about the shadow boundaries, the fields associated with the incident rays, the reflected rays and the diffracted rays are all distinct and can be computed using the Geometrical Theory of Diffraction (GTD).

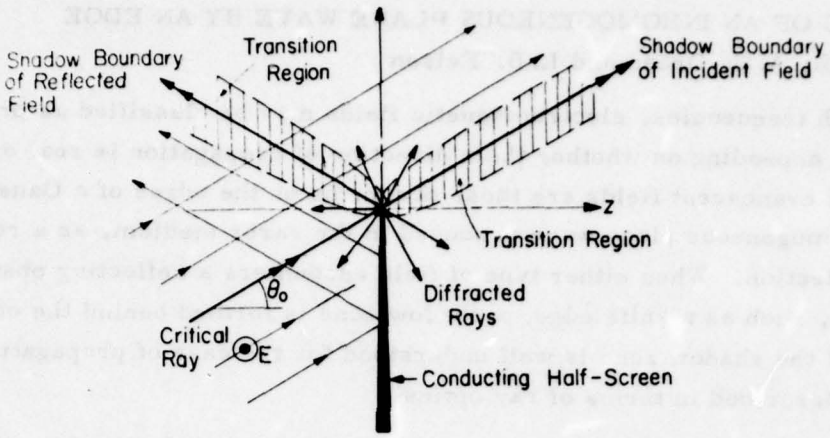


Fig. 1. Shadow formation and diffraction for a homogeneous plane wave incident on a conducting half-screen.

Within the transition regions one cannot find the field using the simple GTD formulas, which are discontinuous or divergent at the shadow boundaries. Instead, modified expressions that give a continuous variation of the field across the boundary must be employed.[3, 4] The failure of the simple GTD formulas reflects the fact that the rays of two different families are nearly parallel in the transition region, and therefore the two families do not have distinct physical properties there. Thus while a precisely defined shadow boundary can be envisioned mathematically, the character of the fields in the transition region is such that the boundary cannot be localized with infinite precision, except in the limit of zero wavelength.

Evanescent fields in a lossless medium, cannot be described in terms of rays whose paths lie in real space,[5-9] even at high frequencies. The direction of propagation of the fields, and hence the angle made with some real coordinate axis, is complex. When such fields are diffracted by an obstacle, the shadow boundary cannot therefore be defined by the critical rays described above. While it seems evident that the incident wave illuminates certain regions and is blocked by the obstacle from reaching other (shadow) regions, it is not a priori evident where the shadow boundary lies. Furthermore, the variation of the fields in the vicinity of the shadow boundary, and hence the character of the shadow, are not clear since the exponential decay of the incident wave may imply that its field is smaller than the diffracted field along the shadow boundary.

The problem of locating the shadow boundary has been previously recognized,[5] but has not been settled conclusively. For a smoothly curved cylindrical obstacle, a

study of the creeping wave behavior revealed a shift of the boundary from its geometric optical location when the incident field is evanescent.[10] While the physical process associated with wave propagation in a complex direction was studied,[9] it was not used to predict the location of shadow boundaries. In treating edge diffraction of a Gaussian beam using complex rays, Otis[11] somewhat arbitrarily defined the shadow boundary by taking the magnitude of the complex transverse coordinate of the complex ray passing through the edge, but did not investigate the character of shadow formation. This prescription for the shadow boundary is not consistent with that obtained here, but the error that is introduced in the fields by its use is smaller than the amplitude of the diffracted field.

B. Evaluation of the Field and the Shadow Boundaries

In order to gain detailed insight into the location and character of shadow boundaries for inhomogeneous waves we have returned to the problem of evanescent plane wave diffraction by a conducting half-screen. The geometrical configuration is as shown in Fig. 2, where the angle $\theta_o = u + iv$. A harmonic plane wave, with time dependence exp $(-i\omega t)$ implied, is assumed to be incident in the x-z plane on a perfectly conducting half-screen lying in the half-plane $x < 0$, $z = 0$, as shown in Figure 2. For

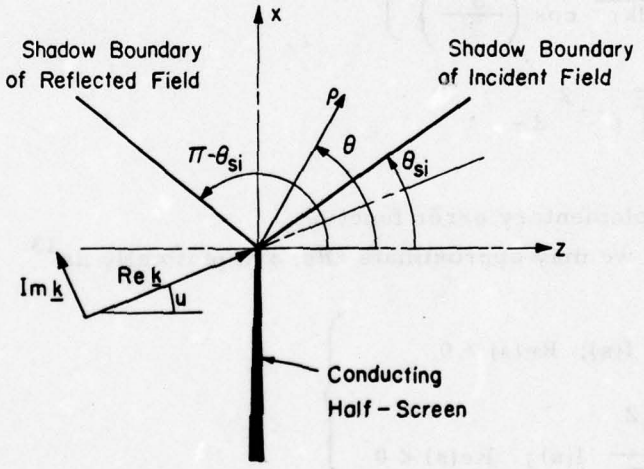


Fig. 2. Shadow boundaries for an inhomogeneous plane wave incident on a conducting half-screen at a complex angle $\theta_o = u + iv$ for the case $v > 0$. The phase paths[o] [10] for the incident field is the family of lines parallel to $\theta = u$.

simplicity, the electric field is assumed to be polarized along y. The variation of the electric field is given by

$$E_{inc}(x, z) = \exp[\, ik\rho\cos(\theta-\theta_o)] = \exp[i\underline{k}\cdot\underline{\rho}]\,, \tag{1}$$

where $k = \omega\sqrt{\epsilon\mu}$ is the wavenumber of the medium and the cylindrical coordinates (ρ, θ) are as indicated in Figure 2. In Eq. (1), θ_o is the angle that the wave vector

$$\underline{k} = \underline{x}_o\, k \sin\theta_o + \underline{z}_o\, k\cos\theta_o \tag{2}$$

makes with the z axis. For an inhomogeneous plane wave in a lossless medium (k real), $\theta_o = u + iv$ is complex and correspondingly, $\underline{k} = \mathrm{Re}\,\underline{k} + i\,\mathrm{Im}\,\underline{k}$ is complex. The vectors $\mathrm{Re}\,\underline{k}$ and $\mathrm{Im}\,\underline{k}$ are perpendicular to one another as indicated in Fig. 2, and $\mathrm{Im}\,\underline{k}$ points in the direction of maximum rate of decay of the fields.

As a result of reflection from the screen and diffraction at its edge, the total electric field $E(\rho, \theta)$, which includes the incident plane wave, is found to be [12]

$$E(\rho, \theta) = \exp[ik\rho\cos(\theta-\theta_o)]\, Q(p) - \exp[-ik\rho\cos(\theta-\theta_o)]\, Q(q). \tag{3}$$

Here

$$\left.\begin{aligned} p &= e^{-i\pi/4}\sqrt{2k\rho}\,\sin\left(\frac{\theta_o-\theta}{2}\right) \\ q &= e^{-i\pi/4}\sqrt{2k\rho}\,\cos\left(\frac{\theta_o+\theta}{2}\right) \end{aligned}\right\} \tag{4}$$

and

$$Q(s) = \frac{1}{\sqrt{\pi}}\int_s^\infty e^{-\tau^2}\, d\tau \tag{5}$$

is one half the complementary error function.

For $|s| \gg 1$, we may approximate $Q(s)$ aymptotically as [13]

$$\left.\begin{aligned} Q(s) &\sim \frac{e^{-s^2}}{2\sqrt{\pi s}}\, I(s);\quad \mathrm{Re}(s) > 0 \\ Q(s) &\sim 1 + \frac{e^{-s^2}}{2\sqrt{\pi}\, s}\, I(s);\quad \mathrm{Re}(s) < 0 \end{aligned}\right\} \tag{6}$$

with

$$I(s) = 1 + \sum_{n=1}^{N} \frac{(1\cdot 3\cdot 5\ldots 2n-1)}{(-2)^n\, s^{2n}} \tag{7}$$

The series in Eq. (7) is asymptotic in the sense that the terms initially decrease in magnitude for fixed large s, but beyond a certain value of n they increase in magnitude. While taking an infinite number of terms would cause I(s) and thereby Q(s) to diverge, a limited number of terms will give a good approximation to Q(s). Since the error is of the order of the first term neglected, [14] the most accurate sum for Q(s) is obtained by cutting off at a value N such that the n = N + 1 term is minimum. Because of the variation of the terms with n, as cited above, N can be found from the condition that the n = N and n = N + 1 terms be the two most nearly equal in magnitude. This definition of N is implied in Eq. (7), recognizing that N will be a function of the coordinates (ρ, θ).

Using Eqs. (7) and (8) in Eq. (4), the total field can be separated into a geometrical optics component E_{GO} and a diffracted component E_D.
The geometrical optical component is

$$E_{GO} = \exp[\, ik\rho\cos(\theta-\theta_o)]\; U(\theta-\theta_{si}) - \exp[\, -ik\rho\cos(\theta+\theta_o)]\; U(\theta-\theta_{sr}). \tag{8}$$

Here $U(\tau)$ is the unit step function and the angles θ_{si} and θ_{sr} are found from the conditions Re p = 0 and Re q = 0, respectively. The first term in Eq. (8) represents the incident plane wave, which illuminates the region $3\pi/2 > \theta > \theta_{si}$, but does not reach the region $\theta_{si} > \theta > -\pi/2$. The second term is the plane wave reflected from the half-screen, and occupies the region $3\pi/2 > \theta > \theta_{sr}$. It is evident from Eq. (8) that θ_{si} and θ_{sr} give the shadow boundaries of the incident and reflected fields. From Eq. (4) and the conditions cited after Eq. (8) these boundaries are given by

$$\left.\begin{aligned} \theta_{si} &= u + \tan^{-1}(\sinh v) \\ \theta_{sr} &= \pi - \theta_{si} \end{aligned}\right\} . \tag{9}$$

For the case of a homogeneous plane wave, v = 0 and $u = \theta_o$, so that $\theta_{si} = \theta_o$ and $\theta_{sr} = \pi - \theta_o$ from Eq. (9) and the shadow boundaries are obtained from the critical ray, as in Figure 1. When θ_o is complex, Re $\underline{k}$ makes an angle u with the z-axis, as in Figure 2. For v > 0, Im $\underline{k}$ makes an angle u with the positive x-axis, as shown in Figure 2. In this case the shadow boundary $\theta = \theta_{si}$ for the incident plane wave lies to the right of the screen and above the radial line $\theta = u$. Because the field decays in the direction of Im $\underline{k}$, and since $\theta_{si} > u$, the incident field along its shadow boundary decays away from the edge and therefore becomes exponentially small compared to the diffracted field, which is given below. For v < 0, Im $\underline{k}$ has a negative x component and $\theta_{si} < u$, so that for this case also, the incident field is exponentially small along its shadow boundary. Similar statements apply to the variation of the reflected field along its shadow boundary θ_{sr}.

The diffracted field for $|p| >> 1$ and $|q| >> 1$ is obtained from Eqs. (3) and (6) as

$$E_D = \frac{e^{i(k\rho + \pi/4)}}{2\sqrt{2\pi k\rho}} \left[\frac{I(p)}{\sin\left(\frac{\theta_o - \theta}{2}\right)} - \frac{I(q)}{\cos\left(\frac{\theta_o + \theta}{2}\right)} \right] . \tag{10}$$

For a homogeneous incident plane wave, θ_o is real and E_D as given in [10] is infinite along the shadow boundaries $\theta = \theta_o$, $\pi - \theta_o$. This behavior results from the fact that $|p|$ or $|q|$ vanishes on the shadow boundary so that the asymptotic expansion is not valid. In the case of an inhomogeneous plane wave, θ_o is complex so that $\sin\left(\frac{\theta_o - \theta}{2}\right)$ or $\cos\left(\frac{\theta_o + \theta}{2}\right)$ do not vanish along the shadow boundary. Thus for $k\rho$ sufficiently large, Eq. (10) will be valid on the shadow boundaries, while if $k\rho$ is small enough, the condition $|p| >> 1$ or $|q| >> 1$ will not be satisfied there, and hence Eq. (10) cannot be used. In the next section we examine, for θ_o complex, the regions in space where the conditions for Eq. (10) to hold are satisfied, and we explore the character of the field in the vicinity of the shadow boundaries.

C. Transition Regions and the Character of the Shadow Boundary

Based on the asymptotic approximation of the complementary error function, we have defined the shadow boundaries produced by the edge of the screen. The nature of the shadowing effect across the boundaries remains to be determined. In order for the asymptotic approximation leading to Eqs. (8) and (10) to be valid, it is necessary that $|p| >> 1$ and $|q| >> 1$, the first condition being significant for the shadow boundary θ_{si} of the incident field and the second applying to the shadow boundary $\pi - \theta_{si}$ of the reflected field.

Let A be a large number such that the conditions $|p| >> 1$ and $|q| >> 1$ are satisfied whenever $|p| \geq \sqrt{A}$ and $|q| \geq \sqrt{A}$. The significance of this number will be considered subsequently. For the incident field, the condition $|p| \geq \sqrt{A}$ will be satisfied at points (ρ, θ) outside of a transition region whose boundary is given by

$$\rho\,[\cosh v - \cos(\theta - u)] = A/k . \tag{11}$$

In the case of a homogeneous plane wave with $v = 0$, the curve given by Eq. (11) is a parabola with focus at the edge $\rho = 0$ and centered about the shadow boundary $\theta = \theta_{si} = \theta_o$.

In the case of an inhomogeneous plane wave, the curve described by Eq. (11) is an elipse with one focus at the edge $\rho = 0$ and major axis oriented along the radial line $\theta = u$, as indicated in Figure 3. The length L and width W of the ellipse are

$$\left.\begin{aligned} L &= 2A \cosh v/(k \sinh^2 v) \\ W &= 2A/k \sinh v \end{aligned}\right\} . \tag{12}$$

The shadow boundary θ_{si} intersects one side of the ellipse at its widest point, as in Fig. 3 for the case $v > 0$. For $v < 0$, the boundary θ_{si} intersects the ellipse on the opposite side. Thus the region in which the asymptotic evaluation is not valid lies near, but is not centered on, the shadow boundary. Moreover, it has only finite extent. For example, if $|\text{Im } \underline{k}| = (0.01)|\text{Re } \underline{k}|$ and $u = 0$, then $v \approx 0.01$ and the decay rate in the direction of Im $\underline{k}$ is such that the field decreases by the factor $1/e$ in a distance that is approximately 16λ. Taking $A = 10$, one has $W \approx 320\lambda$ and $L \approx 3 \times 10^4 \lambda$, so that for optical frequencies $L \sim 1.5$ cm, which is significant for laboratory measurements.

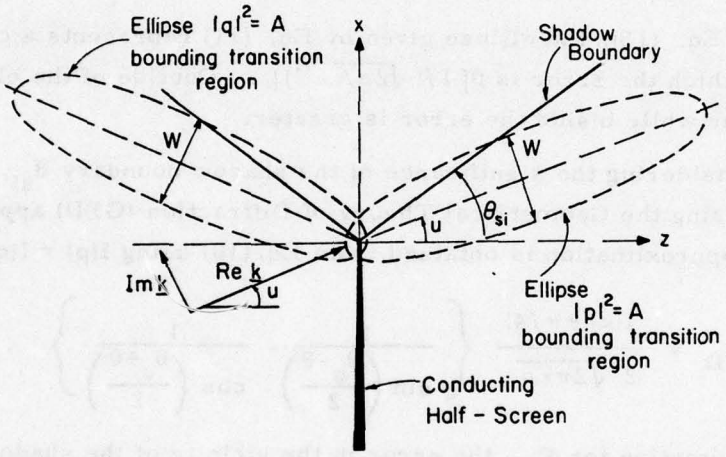


Fig. 3. Transition regions outside of which asymptotic approximation is valid for the case $v > 0$.

To understand the significance of the shadow boundary, especially outside of the transition zone, it is necessary to examine the error involved in using the asymptotic approximation. The error made in using the asymptotic series for I(p) and I(q) in Eq. (10) is of the order of the first term neglected in each series. In the vicinity of the shadow boundary at θ_{si}, the error will predominantly be that associated with the series for I(p). Taking the optimum number of terms N in the series for I(p), as defined after Eq. (7), the error made in using Eq. (10) for the diffracted field will be of order

$$\text{error } (N) = 0 \left\{ \frac{(1.3.5. \ldots 2N+1)}{2\sqrt{2\pi k\rho}\left|\sin\left(\frac{\theta_o - \theta}{2}\right)\right|^{(2^{N+1})}\left|p^{2(N+1)}\right|} \right\} \tag{13}$$

The value of N is a function of $|p|$, and hence of position (ρ, θ), which can be found by requiring the $n = N$ and $n = N + 1$ terms in Eq. (7), to be those two most nearly equal in magnitude. Thus N is the integer closest to the number $\overline{N}$ given by

$$\overline{N} = |p^2| - \frac{1}{2} \quad . \tag{14}$$

Using Eqs. (14) in (13) it is found after some manipulation that

$$\text{error } (N) = 0\left\{\frac{1}{\sqrt{2\pi}} \; \frac{e^{-|p^2|}}{|p|}\right\}$$

$$= 0\left\{\frac{e^{-2k\rho\left|\sin\left(\frac{\theta-\theta_o}{2}\right)\right|^2}}{2\sqrt{\pi k\rho}\left|\sin\left(\frac{\theta-\theta_o}{2}\right)\right|}\right\} \quad . \tag{15}$$

In view of Eq. (15), the ellipse given by Eq. (11) represents a contour of constant error on which the error is $0[1/(\sqrt{2\pi A}\, e^A)]$. Outside of the ellipse, the error has a lower value while inside the error is greater.

Before considering the significance of the shadow boundary θ_{si}, we examine the error made in using the Geometrical Theory of Diffraction (GTD) approximation for E_D. The GTD approximation is obtained from Eq. (10) using $I(p) = I(q) = 1$, so that

$$E_D \approx E_{GTD} = \frac{e^{i(k\rho+\pi/4)}}{2\sqrt{2\pi k\rho}}\left\{\frac{1}{\sin\left(\frac{\theta_o-\theta}{2}\right)} - \frac{1}{\cos\left(\frac{\theta_o+\theta}{2}\right)}\right\} \quad . \tag{16}$$

With this approximation for E_D, the error in the vicinity of the shadow boundary θ_{si} is the order of the field contribution from the $n = 1$ term of $I(p)$. From Eqs. (7) and (10), it is seen that

$$\text{error } (0) = 0\left\{\frac{1}{2\sqrt{\pi}|p^3|}\right\} \quad . \tag{17}$$

Thus, the ellipse of Eq. (11) also represents a contour of constant error for the GTD approximation, Eq. (16) on which the error is $0[1/(2\sqrt{\pi}\, A^{3/2})]$.

The significance of the shadow boundary θ_{si} can be seen by comparing the error, Eq. (15), to the incident field. After some manipulation it is found that

$$\frac{\text{error } (N)}{|E_{inc}|} = 0\left\{\frac{1/(2\sqrt{\pi k\rho})}{\left|\sin\left(\frac{\theta-\theta_o}{2}\right)\right|} \exp\left[-\frac{k\rho}{\cos(\theta_{si}-u)}\{1-\cos(\theta_{si}-\theta)\}\right]\right\} . \tag{18}$$

Outside of the ellipse of Eq. (11) with A = 1, the ratio of Eq.(18) is seen to be less than unity for all θ. For $k\rho >> 1$, the dependence of Eq.(18) on θ is dominated by the exponential term, which is greatest when $\theta = \theta_{si}$. Thus, error (N) is most nearly equal to $|E_{inc}|$, along the radial line $\theta = \theta_{si}$, whereas error (N) decreases in comparison to $|E_{inc}|$ as θ deviates from θ_{si}.

To see the significance of Eq.(18), suppose that we had arbitrarily taken the shadow boundary to be at some angle $\overline{\theta}_{si}$ other than that given by Equation (9). In this case the first term in E_{GO} of Eq. (8) would have as a factor the unit step function $U(\theta - \overline{\theta}_{si})$, instead of $U(\theta - \theta_{si})$. The discontinuity generated in the asymptotic approximation by $U(\theta - \overline{\theta}_{si})$ would, in view of Eq. (18), be larger in comparison to error (N) then if we had taken $\overline{\theta}_{si} = \theta_{si}$. In other words, the discontinuity produced by dropping E_{inc} as the shadow boundary is crossed has the least effect on the overall error when the shadow boundary is taken along the radial line $\theta = \theta_{si}$. Because E_{inc} is exponentially small compared to E_D along the shadow boundary, its omission is referred to in the mathematical literature on asymptotics as the Stokes phenomenon, and the shadow boundary is known as a Stokes line.[15] If one is satisfied with less than the highest accuracy possible and uses the GTD approximation Eq. (16), for the diffracted field, the shadow boundary may be more loosely defined, provided that the magnitude of the incident field along the boundary does not exceed the error in Equation (17). However, the choice $\theta_{si} = u$, as implied in Ref. 10 is not valid since $|E_{inc}| = 1$, and therefore its omission on one side of the radial line $\theta_{si} = u$ would result in an unacceptably large discontinuity.

The exponential factor in Eq. (18) is constant on contours given by

$$k\rho[1 - \cos(\theta_{si} - \theta)] = B, \tag{19}$$

where B is some constant. Expression (19) is the equation of a parabola about $\theta = \theta_{si}$ with focus at the edge. In the limit as $v \to 0$, the ellipse, Eq. (11), which is a contour of constant absolute error, coincides with the parabola, Eq. (19), for B = A. For evanescent fields, the two contours are distinct.

The error associated with the asymptotic approximation of Q(p) in Eq. (3) has the same properties as those described above for Q(p), except that the relevant geometrical shapes are the mirror images in the (x-y) plane of those for Q(p), as in Figure 3. Thus, the contours of constant absolute error are ellipses centered about the line $\theta = \pi - u$. Along the shadow boundary $\theta = \theta_{sr} = \pi - \theta_{si}$, error (N) for Q(q) is largest compared to the magnitude of the reflected field, given by the second exponential in Equation (10). Thus dropping the reflected field across the radial line $\theta = \theta_{sr}$ produces the least discontinuity in the total field, relative to the error inherent in the asymptotic approximation.

D. Summary

This study of high-frequency evanescent plane wave diffraction by a perfectly conducting semi-infinite plane has clarified the location of the shadow boundaries, the transition regions surrounding them, and the mechanism of formation of the shadow. The concepts of illuminated and shadow zones are meaningful only at high frequencies and hence are tied intimately to the asymptotic behavior of the rigorously formulated field. It has been found that, as in the case of homogeneous plane wave diffraction, the high frequency field can be separated into incident, reflected, and edge diffracted constituents. However, sufficiently far from the edge, this quasi-optic description holds for all observation angles, without need of transition functions, since the transition regions, in contrast to the homogeneous wave case, are of finite extent (see Figure 3). The reason for this simplification is the negligible discontinuity in the field values caused by deletion of the evanescent incident field on crossing the shadow boundary. To achieve this behavior, the shadow boundary must be chosen appropriately and related to the error inherent in the asymptotic field representation. For minimal error, this has led to an identification of the shadow boundaries with the Stokes lines in the asymptotic field expansion.

When a larger than minimal error is acceptable in the asymptotic field formulation, as for example by retaining only the leading term in the asymptotic expansion of the edge diffracted field, the location of the shadow boundary becomes less precise. A range of contours near the optimal location exists whenever the incident field is smaller than the permitted error. Yet it is important to observe that the intuitively appealing incident phase path $\theta = u$ that grazes the edge does not meet this requirement and is therefore unacceptable. The same behavior is found to occur when the incident field is a Gaussian beam.[16]

Joint Services Technical Advisory Committee
F44620-74-C-0056 H. L. Bertoni, A. C. Green and L. B. Felsen

U.S. Army Research Office
DAHC04-75-G-0152

REFERENCES

1. G. A. Deschamps, "Ray Techniques in Electromagnetics," Proc. IEEE, Vol. 60, pp. 1022-1035 (1972).

2. J. B. Keller, "Geometrical Theory of Diffraction," J. Opt. Soc. Am., Vol. 52, pp. 116-130 (1962).

3. R. G. Kouyoumjian and P. H. Pathak, "A Uniform Theory of Diffraction for an Edge in a Perfectly Conducting Surface," Proc. IEEE, Vol. 62, pp. 1448-1461 (1974).

4. D. S. Ahluwalia, R. M. Lewis and J. Boersma, "Uniform Asymptotic Theory of Diffraction by a Plane Screen," SIAM J. Appl. Math., Vol. 16, pp. 783-807 (1968).

5. W. Y. D. Wang and G. A. Deschamps, "Application of Complex Ray Tracing to Scattering Problems," Proc. IEEE, Vol. 62, pp. 1541-1551 (1974).

6. K. G. Budden and P. D. Terry, "Radio Ray Tracing in Complex Space," Proc. Roy. Soc., London, Ser. A, Vol. 321, pp. 275-301 (1971).

7. Y. A. Kravtsov, "Complex Rays and Complex Caustics," Bell. Lab. Translation TR-69-109 (1969) of paper presented at 4th All-Union Symp. on Diffraction of Waves, Kharkov, USSR (1967).

8. L. B. Felsen, "Evanescent Waves," J. Opt. Soc. Am., Vol. 66, pp. 751-760 (1976).

9. H. L Bertoni, L. B. Felsen and A. Hessel, "Local Properties of Radiation in Lossy Media," IEEE Trans. Antennas and Propag., Vol. AP-19, pp. 226-237 (1971).

10. S. Choudhary and L. B. Felsen, "Asymptotic Theory for Inhomogeneous Waves," IEEE Trans. Antennas and Propag., Vol. AP-21, pp. 827-842 (1973).

11. G. Otis, "Application of the Boundary-Diffraction-Wave Theory to Gaussian Beams," J. Opt. Soc. Am., Vol. 64, pp. 1545-1550 (1974).

12. L. B. Felsen and N. Marcuvitz, "Radiation and Scattering of Waves," (Englewood Cliffs, N. J.: Prentice Hall, Inc., 1973) p. 673.

13. a. M. Abramowitz and I. A. Stegun, "Handbook of Mathematical Functions," (New York: Dover Pub, 1964) pp. 297-298 gives the algebraic form in Equation (7).

b. H. Jeffreys, "Asymptotic Approximations," (Oxford: Clarendon Press, 1962) p. 42, discusses the range of s over which to include the unit term in Equation (7).

14. H. Jeffreys, op. cit., pp. 117-122.

15. H. Jeffreys, op. cit., pp. 35, 42.

16. H. L. Bertoni, L. B. Felsen and A. Green, "Shadowing of a Gaussian Beam by an Edge," in preparation.

RADIO-WAVE PROPAGATION ALONG MIXED PATHS IN FOREST ENVIRONMENTS

T. Tamir

The propagation of radio waves in the presence of wooded areas has been the subject of intensive studies,[1] but most investigations have dealt with communication paths located entirely within or close to the vegetation. In particular, numerous field-strength measurements have recently been obtained along paths that connect two points inside a forest, but fewer such measurements were taken between a point inside the vegetation and another point outside the wooded area. Similarly, theoretical studies have concentrated on propagation paths that lie entirely inside the vegetation region and less attention has been paid to paths that lie partly outside the forest.

The aim of the present report is to take a closer look at the propagation mechanism for mixed paths in forest environments and to present analytical results for the relevant field amplitudes. An important feature of this work is to show that these analytical results can be described in terms of ray paths which permit the calculation of radio losses in a large class of forest environments. The extension of this ray approach to more complex situations is straightforward. As an example, the ray representation is discussed in the context of a propagation path that traverses through two wooded regions which are separated by a region containing no vegetation.

In agreement with previous work,[1-4] the idealized model for mixed-path considerations in a forest environment is given by the configuration shown in Figure 1. A forest

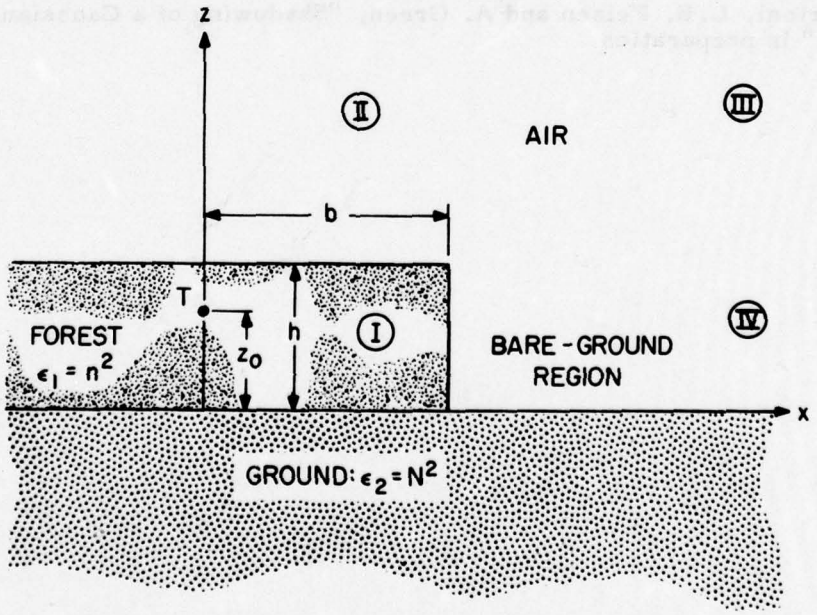


Fig. 1. Geometry of a forest layer adjacent to a bare ground region.

layer having an effective tree height h is shown to the left of the x = b plane, with the space to the right of that plane being a bare ground region. Within the frequency range of (approximately) 2-200 MHz, the vegetation layer is assumed to be a homogeneous refractive medium with a relative permittivity $\epsilon_1 = n^2$, where n is the average index of the vegetation. Similarly, the relative permittivity of the ground is denoted by $\epsilon_2 = N^2$.

To arrive at simple field considerations, we assume that a dipole antenna is placed at the point $T = T(0, 0, z_o)$ where $z_o \leq h$ is inside the forest. Both vertical and horizontal polarization will be examined but, for simplicity, we consider only the field detected at a point $R = R(x, 0, z)$ by a dipole oriented for maximum reception of the signal.

An important feature of the forest-layer geometry is that, depending on the particular location of the receiving point R, four different canonical situations (or propagation regimes) exist, each of which is characterized by a different type of wave mechanism and therefore by a different expression for the electromagnetic field. As indicated in Fig. 1 by Roman numerals, these various regimes correspond to the receiving point being located as follows: (I) inside the vegetation layer, (II) in the air region above the vegetation layer, (III) at a relatively high altitude in the air above the bare ground region, and (IV) at a relatively low height above the bare-ground region.

A. The Field Inside the Forest (Region I)

Because the wave reflections are negligible at the forest-air interface, the forest boundary at x = b may be neglected for reception points R(x, 0, z) located inside the vegetation layer. In this case, the forest may be assumed to extend also over the entire region x > b, as shown in Figure 2.

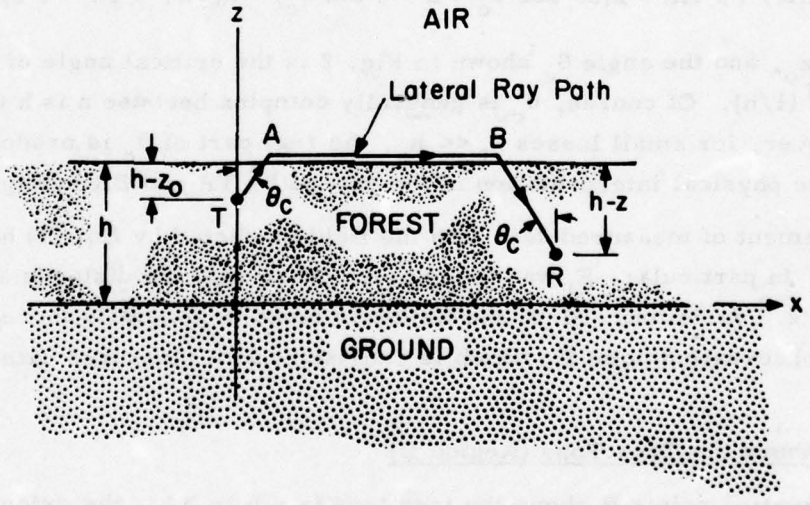


Fig. 2. Ray path TABR for the lateral wave inside the forest layer (Region I).

When the forest layer covers the entire ground plane, the field problem reduces to a situation whose solution is already available in the literature.[2-4] To establish

notation, we repeat here the pertinent results for the electric field E detected at points inside the forest ($x \leq b$, $0 \leq x \leq h$). For large distances $|x|$ and a time dependence $\exp(-i\omega t)$, which is suppressed, a saddle-point evaluation of $E = E_I$ (in Region I) yields

$$E_I = \frac{60\,Ia}{n^2 - 1} \frac{e^{ik[\,|x| + \sqrt{n^2 - 1}\,(2h - z - z_o)]}}{x^2} F(90^o, z)\, F(90^o, z_o) \quad , \tag{1}$$

where Ia is the dipole moment, $k = 2\pi/\lambda$ is the plane-wave propagation factor in air and

$$F(\theta, z) = \frac{1 + B(\theta, z)}{1 - B(\theta, h)} \quad , \tag{2}$$

$$B(\theta, z) = \Gamma(\theta)\, e^{2ikz\sqrt{n^2 - 1}} \tag{3}$$

$$\Gamma(\theta) = \frac{M(n^2 - \sin^2\theta)^{1/2} - m(N^2 - \sin^2\theta)^{1/2}}{M(n^2 - \sin^2\theta)^{1/2} + m(N^2 - \sin^2\theta)^{1/2}} \quad . \tag{4}$$

$$M = m = 1 \quad , \quad \text{for horizontal polarization} \tag{5}$$

$$M = N^2;\ m = n^2 \quad , \quad \text{for vertical polarization} \quad .$$

The field given by E_I is that of a lateral wave, which follows the path TABR in Fig. 2, as can be verified by noting that the exponent in Eq. (1) represents an electrical length given by

$$kn(\overline{TA} + \overline{BR}) + k\,\overline{AB} = k(ns \sec\theta_c + x - s \tan\theta_c) = k(|x| + \sqrt{n^2 - 1}\, s) \quad ,$$

where $x = 2h - z - z_o$, and the angle θ_c shown in Fig. 2 is the critical angle of total reflection: $\theta_c = \sin^{-1}(1/n)$. Of course, θ_c is generally complex because n is a complex quantity. However, for small losses $n_i \ll n_r$, the real part of θ_c is predominant and it then yields the physical interpretation of the ray-paths TA and BR in Figure 2.

The agreement of measured data with the field predicted by Eq. (1) has been well documented.[1-3] In particular, E_I was found to decrease with the distance squared, as required by the x^{-2} variation. Also, the field was found to generally increase with the height z of the observation point R, which is consistent with measured data.

B. The Field Above the Tree Tops (Region II)

For observation points R above the tree tops ($x \leq b$, $z \geq h$), the extended slab geometry used above also holds because the forest boundary at $x = b$ has a negligible effect on the wave reaching the receiving dipole. In this case, a first-order saddle-

point evaluation for large distances r >> h yields a geometric-optical result

$$E_{II}^{(o)} = \frac{120\pi\, i\, Ia\, p}{m\cos\theta + \sqrt{n^2 - \sin^2\theta}} \frac{1}{1 - \gamma(\theta)\, B(\theta, h)} \frac{e^{ik[r + \sqrt{n^2 - \sin^2\theta}\,(h - z_o)]}}{\lambda r} \cos\theta \qquad (6)$$

where r and θ are indicated in Fig. 3, and

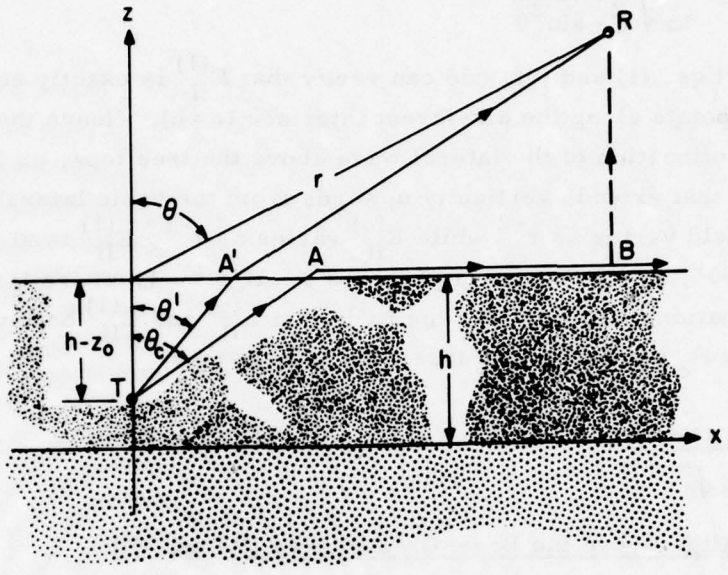


Fig. 3. Ray paths TA'R for the reflected wave and TABR for the lateral wave above the tree tops (Region II). Because of far-field assumptions, the radius vector r is essentially parallel to the refracted ray A'R.

$$\gamma(\theta) = \frac{\sqrt{n^2 - \sin^2\theta} - m\cos\theta}{\sqrt{n^2 - \sin^2\theta} + m\cos\theta} \qquad (7)$$

$$p = \begin{cases} 1, & \text{for horizontal polarization} \\ \sin^2\theta, & \text{for vertical polarization} \end{cases} \qquad (8)$$

The field $E_{II}^{(o)}$ represents a refracted wave, as shown by the ray TA'R in Figure 3. The geometric-optical character of this field is evidenced by the spherical-wave dependence r^{-1} and by the exponent in Eq. (6), which yields an electrical length

$$kn \cdot \overline{TA'} + k \cdot \overline{A'R} = k[n(h - z_o)\sec\theta' + r - (h - z_o)\tan\theta'\sin\theta]$$

$$= k[r + \sqrt{n^2 - \sin^2\theta}\,(h - z_o)] \quad ,$$

where θ', shown in Fig. 3, is related to θ via Snell's condition $\sin\theta = n \sin\theta'$.

The field $E_{II}^{(o)}$ is, however, only a partial component of the total field E_{II}. In particular, it is noted that $E_{II}^{(o)}$ is proportional to $\cos\theta$ and therefore vanishes at the forest-air boundary ($z = h$, $\theta = 90^{o}$). For points in the vicinity of $\theta = 90^{o}$, a second-order saddle-point contribution yields another field component which is given by

$$E_{II}^{(1)} = E_{II}^{(o)} \frac{im \sin\theta}{kr\sqrt{n^2 - \sin^2\theta}} F(\theta, h) \tan\theta \quad . \tag{9}$$

By comparing Eqs. (1) and (9), one can verify that $E_{II}^{(1)}$ is exactly equal to the lateral wave E_I at all points along the air-forest interface ($z = h$). Hence the field $E_{II}^{(1)}$ is identified as the continuation of the lateral wave above the tree tops, as indicated by the dashed ray BR that extends vertically upwards from the basic lateral-ray path TAB. Because this field varies as r^{-2} while $E_{II}^{(o)}$ varies as r^{-1}, $E_{II}^{(1)}$ is significant only for angles θ near 90°, in which case $\cot\theta \simeq \cos\theta = H/r \ll 1$, where $H = z - h$. Inserting these approximations into Eq. (9), one finds that $E_{II}^{(o)}$ and $E_{II}^{(1)}$ become equal in magnitude at a height H_c above the tree tops, which is given by

$$\frac{H_c}{\lambda} \simeq \left| \frac{m F(90^{o}, h)}{2\pi\sqrt{n^2 - 1}} \right| \quad . \tag{10}$$

C. The Field High Above the Bare-Ground Area (Region III)

Because the field at heights $z \geq h + H_c$ above the tree tops is given essentially by a refracted ray, as shown above, it follows that this ray is responsible for the field also at points of a ray path that penetrates into the region above the bare ground area. This argument is illustrated in Fig. 4, where the refracted ray TA'R is extended in Region III. Consequently, the field E_{III} in this region is given simply by

$$E_{III} = E_{II}^{(o)} \quad , \tag{11}$$

except that now the distance r that enters into the expression for E_{III} involves values such that $x > b$.

D. The Field at Low Heights Above the Bare-Ground Area (Region IV)

For observation points R near or below the limit ray TCL shown in Fig. 4, the field cannot be evaluated by directly extending a ray obtained from a previous result. This happens because Region IV is strictly a shadow domain with respect to geometric-optical ray contributions from the transmitter at T; hence the signal reaches into this region only by a diffraction process. However, since the forest layer acts as a "weak" dielectric (with $n \simeq 1$), a rather simple argument can provide a good estimate of the field amplitude in Region IV.

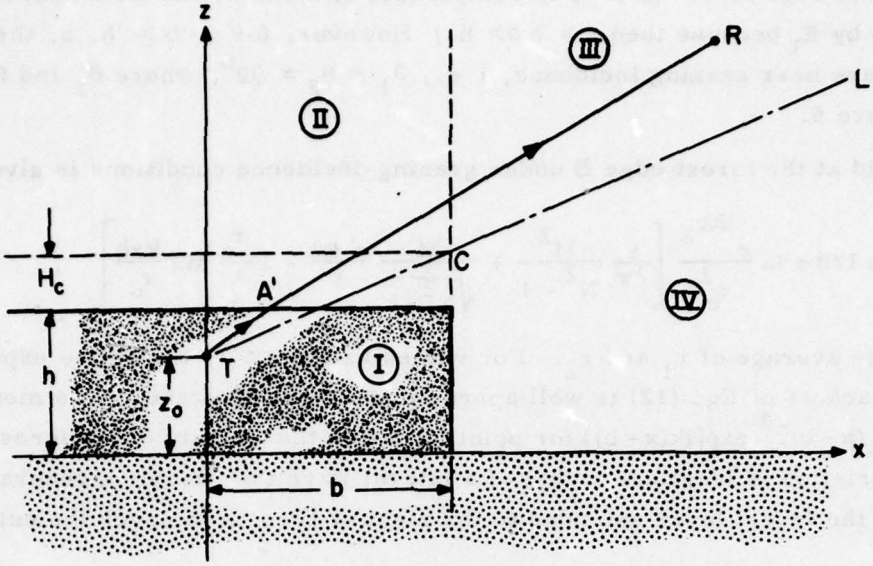


Fig. 4. Extension of the refracted ray TA'R into the region high above the bare-ground area (Region III). The limit ray TCL indicates the approximate location of the boundary between Regions III and IV.

To establish this field evaluation, it is convenient to invert the problem by considering the field at point T produced by a dipole located at point R. Referring to Fig. 5, the field in the vicinity of the forest edge at B is then due essentially to a direct ray

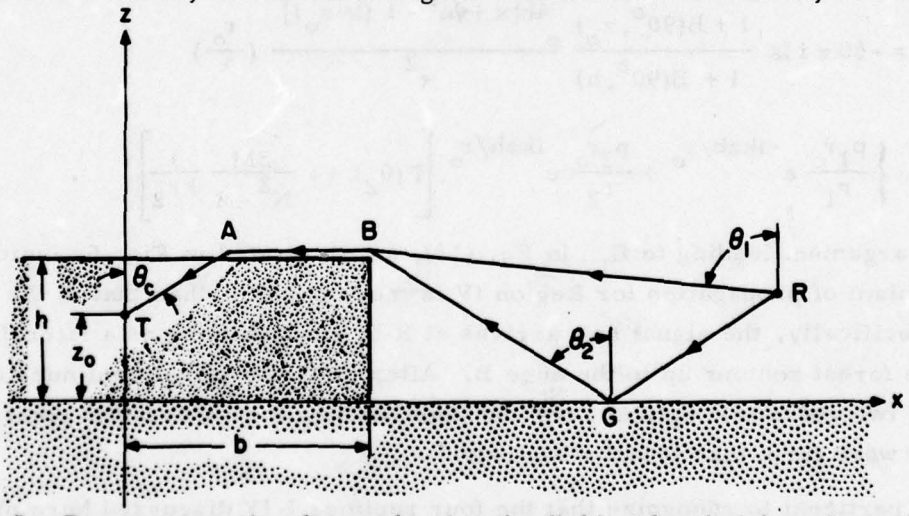


Fig. 5. Ray construction for evaluating the field at low heights above the bare ground area (Region IV). The path of the signal from T to R is obtained by reversing the directions of the arrows.

RB and a ground-reflected ray RGB. In the present case, the situation of practical interest is when both h and z are much smaller than the distance x - b between the point

R and the forest edge at B. (If x - b is comparable to h and z, the field at R is well approximated by E_I because then $x \simeq b \gg h$.) However, for $x - b \gg h, z$, the two rays BR and BGR are near grazing incidence, i.e., $\theta_1 \simeq \theta_2 \simeq 90^\circ$, where θ_1 and θ_2 are shown in Figure 5.

The field at the forest edge B under grazing-incidence conditions is given by

$$E_B \simeq -i\,120\pi\,Ia\,\frac{e^{ikr_o}}{r_o^2}\left[\frac{i}{2\pi}\,\frac{M^2}{N^2-1}+\frac{M}{\sqrt{N^2-1}}\,\frac{z+h}{\lambda}-i\,\frac{r_o}{\lambda}\sin\frac{kzh}{r_o}\right], \tag{12}$$

where r_o is the average of r_1 and r_2. For values $kzh/r_o \ll 1$, the entire expression in the square brackets of Eq. (12) is well approximated by a constant. This means that E_B varies as $(x-b)^{-2}\exp[ik(x-b)]$ for points $z_o = h$ in the vicinity of the forest edge at B. Such a variation is identical to that of a lateral wave that would be generated by the dipole at R if the forest layer extended to the right of B so as to cover the entire bare ground area.

Because the field at, and to the left of, the forest edge B matches the field of a lateral wave, analytical continuation arguments stipulate that the field at any point T inside the forest is found by simply extending the "lateral wave" from B to T. By noting the lateral-wave behavior expressed in Eq. (1), and using reciprocity arguments, Eq. (12) leads to

$$E_{IV} = -60\pi\,i\,Ia\,\frac{1+B(90^\circ,z_o)}{1+B(90^\circ,h)}\,\frac{e^{ik[x+\sqrt{n^2-1}\,(h-z_o)]}}{x^2}\,\left(\frac{r_o}{\lambda}\right)$$

$$\cdot\left\{\frac{p_1 r_o}{r_1}\,e^{-ikzh/r_o}+\frac{p_2 r_o}{r_2}\,e^{ikzh/r_o}\left[\Gamma(\theta_2)+i\,\frac{2M}{N^2-1}\,\frac{1}{kr_2}\right]\right\}. \tag{13}$$

The argument leading to E_{IV} in Eq. (13), as illustrated in Fig. 5, indicates that the mechanism of propagation for Region IV is more complex than that in the other regions. Specifically, the signal now arrives at R by first starting as a lateral wave that follows the forest contour up to the edge B. After that, the signal continues as a combination of two geometric optical rays and thus reaches the observation point at R via a different wave form than it had initially.

It is pertinent to recognize that the four regimes I-IV discussed here play the role of canonical forms, or building blocks, which can be applied to construct the field in more complicated situations. As an example, Fig. 6 shows two separate forest layers of different heights, with a bare-ground region in between. For an antenna located, say, at $T(0, z_o)$, the field at any observation point R(x, z) may be found by using the

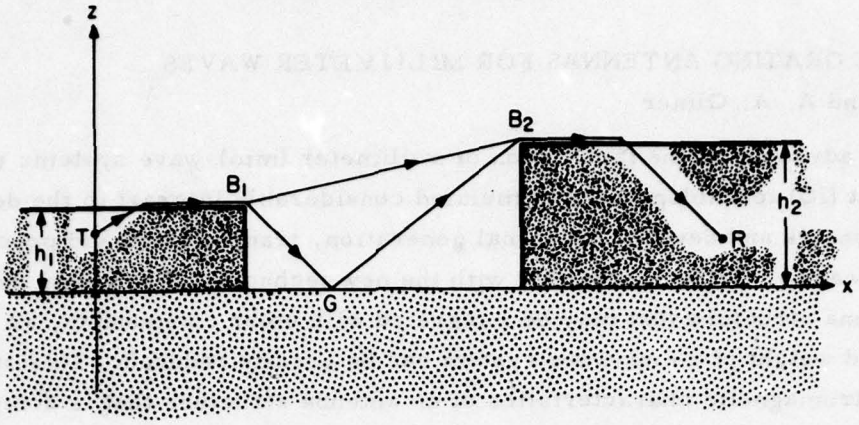


Fig. 6. Ray construction for evaluating the field in a situation containing two forest regions separated by a large clearing.

arguments and results presented here because the field above the bare-ground region is only slightly affected by the presence of the second forest (extending along $x > b_2$). Hence the field is then given by either E_{III} or E_{IV}, depending on the height z. However, inside the second forest (i.e., for $x > b_2$ and $z < h_2$) the field may be obtained by combining three ray constructions in tandem along the lines suggested by Fig. 5, except that now point B_1 plays the role of point R in Figure 5.

Consequently, the field in Fig. 6 is found by starting with a lateral wave from T to B_1, followed by direct and reflected rays connecting B_1 to B_2, after which B_2 is connected to R by another lateral wave. For composite configurations involving more than two forest regions, the field can be similarly tracked from point to point by using the canonical ray construction discussed here. More details on these and other related questions are given in a recent publication.[5]

U.S. Army Electronics Command T. Tamir

Army Land Warfare Laboratory
Aberdeen Proving Ground, Maryland

REFERENCES

1. J.R. Wait, R.H. Ott and T. Telfer (editors), "Workshop on Radio Systems in Forested and/or Vegetated Environments," U.S. Army Communications Command, Fort Huachuca, Ariz., Tech. Rept. No. ACC-ACO-1-74 (February 1974). For background prior to 1974 and detailed bibliography, refer therein to paper I-B: T. Tamir, "Lateral Wave Applications to Radio Systems."

2. T. Tamir, "On Radio-Wave Propagation in Forest Environments," IEEE Trans. Ant. Prop., Vol. AP-15, pp. 806-817 (November 1967).

3. D. Dence and T. Tamir, "Radio Loss of Lateral Waves in Forest Environments," Radio Science, Vol. 4, pp. 307-318 (April 1969).

4. S. Rosenbaum and L.W. Bowles, "Clutter Return from Vegetated Areas," IEEE Trans. Ant. Prop., Vol. AP-22, pp. 227-236 (March 1974).

5. T. Tamir, "Radio-Wave Propagation Along Mixed Paths in Forest Environments," IEEE Trans. Ant. Prop., Vol. AP-25, pp. 471-477 (July 1977).

DIELECTRIC GRATING ANTENNAS FOR MILLIMETER WAVES

S. T. Peng and A. A. Oliner

Recent advances in the fabrication of millimeter (mm)-wave systems using integrated circuit (IC) technology have stimulated considerable interest in the development of new components and devices for signal generation, transmission and processing involving mm-waves. To be compatible with the new technology, it is very desirable to have an antenna structure that may be integrated with other components, so that the cost, size and weight of a mm-wave system can be greatly reduced. This study examines the electromagnetic characteristics of an antenna structure that is compatible with the IC technology for mm-wave applications.

The antenna structure investigated here is a dielectric grating in conjunction with an open dielectric waveguide, which belongs to the general class of traveling wave antennas. Such a composite waveguide structure falls into the general class of periodic dielectric waveguides because it may also serve, with proper designs, many other functions, such as filtering and distributed feedback of electromagnetic waves. As a rigorous electromagnetic boundary value problem, the periodic dielectric waveguide structure has been extensively analyzed in the past, mostly in the context of integrated optics applications.[1-4] The purpose of this work, however, is to investigate the periodic dielectric waveguide with physical and structural parameters that are pertinent to applications in the mm-wave frequency range and with particular attention directed toward the radiation characteristics of such a structure as an antenna.

The requirements on grating antennas for mm-wave applications differ from those for optical couplers in many respects. Because of the longer wavelength, the limitation on the size of grating antennas is very severe. For example, at the frequency of 100 GHz, an antenna 3 inches in length permits only about 25 wavelengths in free space. To radiate almost all the incoming energy within 25 wavelengths (in contrast to more than 1000 wavelengths for optical couplers) the periodic waveguide requires a very large leakage constant. Fortunately, the high dielectric constants of mm-wave materials are very helpful in achieving a large leakage constant, but they may produce numerical difficulties in an exact analysis of the antenna. Also, the previously-developed perturbation analysis that was based on the small difference between the dielectric constants of the grating and the surrounding medium does not yield sufficiently accurate results, and a new approximation technique has to be introduced to obtain better analytic results for dielectric grating antennas of high dielectric constant.

The configuration of a dielectric grating antenna structure is shown in Figure 1. Such an electromagnetic structure has been previously analyzed in the context of optical

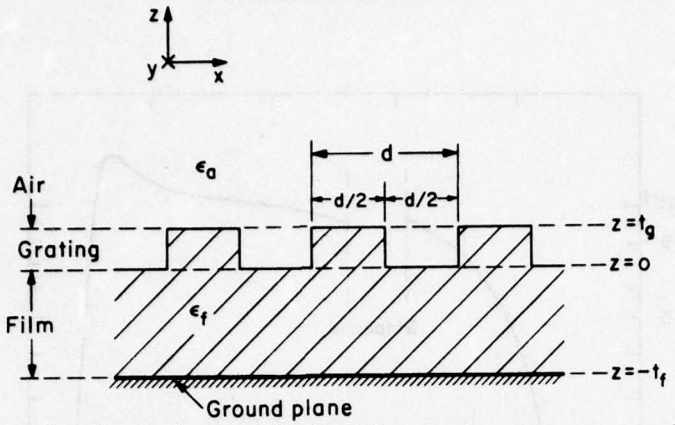


Fig. 1. Configuration of dielectric grating antennas suitable for millimeter wave applications.

periodic couplers[1-4] and many of the radiation characteristics of the structure are known. For example, it was found that the radiation is strongest when the groove width is nearly equal to the tooth width. For this reason, throughout this work we shall choose the width of the grooves to be equal to that of the teeth.

For a given dielectric material, the antenna structure is characterized by the three parameters: the thickness of the uniform film, t_f, the thickness of the grating region, t_g, and the period of the grating, d. In addition, the frequency of the source is characterized by the wavelength, λ. We have determined the effect of each of the four parameters on the radiation characteristics of the antenna with three different dielectric materials: silicon, $\epsilon_f = 12$, Aluminum oxide, $\epsilon_f = 10$ and Boronitride, $\epsilon_f = 4$.

For the three materials under consideration, we have determined that $t_f < 0.25\lambda$ will ensure single mode operation. The radiation rates are found to be sufficiently large over a wide range of radiating angles (from -70° to 70°) for silicon and aluminum oxide antennas, but not for boronitride antennas. As an example, the leakage constant of a silicon grating antenna is shown in Fig. 2 for the range of periodicity lengths d such that only a single beam radiates. The angle of radiation of the beam is directly related to the value of d/λ; on Fig. 2 some selected angles are also indicated. For instance, at $d/\lambda = 0.36$, Fig. 2 shows that $\alpha\lambda \approx 0.1$ and the beam is radiating in the forward direction at about 33 degrees from the normal. For an antenna length $\ell = 20\lambda$, the field strength at the exit-end of the antenna will be smaller than that at the entrance end by a factor $e^{-\alpha\ell} = e^{-2} \approx 0.1$. The means that most of the surface-wave energy can be radiated within 20 wavelengths. The leakage constant α can be easily adjusted (within a maximum limit) by changing the grating thickness. A detailed description of the general procedure for the design of grating antennas is in preparation and will be issued as a final report to the ECOM, Ft. Monmouth, N.J., in the near future.

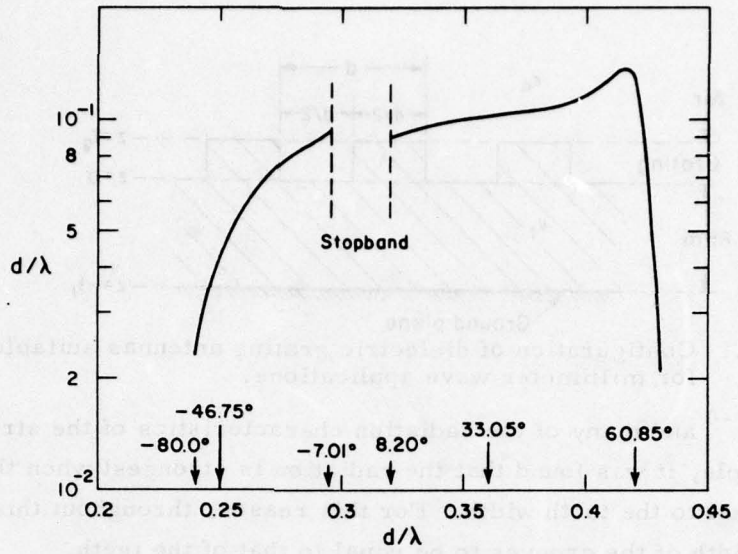


Fig. 2. Leakage constant of silicon dielectric grating antenna vs. the periodicity length. $\varepsilon_f = 12$, $t_f = 0.2\lambda$ and $t_g = 0.05\lambda$.

Joint Services Technical Advisory Committee
F44620-74-C-0056

S. T. Peng

Laboratory Research Cooperative Program
U. S. Army Electronics Command

REFERENCES

1. S. T. Peng, T. Tamir and H. L. Bertoni, "Theory of Periodic Dielectric Waveguides," IEEE Trans. Microwave Theory Tech., Vol. MTT-23, p. 123 (1975).

2. K. Handa, S. T. Peng and T. Tamir, "Improved Perturbation Analysis of Dielectric Gratings," Appl. Phys., Vol. 5, p. 325 (1975).

3. S. T. Peng and T. Tamir, "TM-mode Perturbation Analysis of Dielectric Gratings," Appl. Phys., Vol. 6, p. 35 (1975).

4. T. Tamir and S. T. Peng, "Analysis and Design of Optical Periodic Couplers," Appl. Phys. Accepted for publication.

NOVEL TOTAL-ABSORPTION PHENOMENON OF ENERGY INCIDENT ON PLANAR STRUCTURES

V. Shah and T. Tamir

We show that the energy of a plane wave incident on a layered structure can be totally absorbed if one of the layers has either dielectric or metallic losses. By interpreting this phenomenon in terms of a leaky-wave interaction, we explain similar total absorption effects that have been reported in the context of metallic gratings.

A. Introduction

Recent work on metallic gratings[1-5] has shown that the energy of an incident plane wave can be totally absorbed by the ohmic losses of the metal. In such cases, the reflected and diffracted fields are suppressed, thus obtaining a Brewster-type of phenomenon. This phenomenon was first studied[1,2] for waves with parallel polarization (TM modes) incident on a metal grating and the strong absorption was associated with the excitation of surface plasmons. However, it was later shown[3,4] that the same behavior can be exhibited by waves with perpendicular polarization (TE modes) incident on a grating covered by a dielectric layer; in this case, the strong absorption was related to the excitation of a surface wave supported by the layer.

While the above deals with optical phenomena which can be regarded in electromagnetic terms, the complete suppression of reflected energy by losses has been known for some time in acoustics. In particular, a plane wave incident from a liquid onto a solid can be totally absorbed if the solid medium is slightly lossy.[6] This phenomenon was shown[7] to be produced by the presence of a leaky Rayleigh wave, which also accounts for the lateral shift (Schoch displacement) of bounded acoustic beams.

By using a non-periodic electromagnetic model, which is analogous to the case of acoustic waves in Ref. 7, we show here that total absorption of an incident plane wave can also occur in optics. We then demonstrate that the strong absorption phenomenon is due to a leaky (surface) wave interaction in the presence of either metallic or dielectric losses. We therefore find that this phenomenon is not necessarily restricted to metallic gratings. Instead, total absorption at certain critical angles is found to occur in a wider class of configurations, of which metallic gratings are only a particular case.

B. Plane Wave Reflection by a Layered Lossy Structure

Consider the layered configuration shown in Fig. 1, where a plane wave having perpendicular polarization is assumed to be incident at an angle θ from the upper medium. A total of four media are involved, which are characterized by the dielectric constants ε_d, ε_a, ε_f and ε_m. Of these, ε_d and ε_a are assumed to be real and thus designate lossless media. On the other hand, either ε_f or ε_m may be taken as complex so

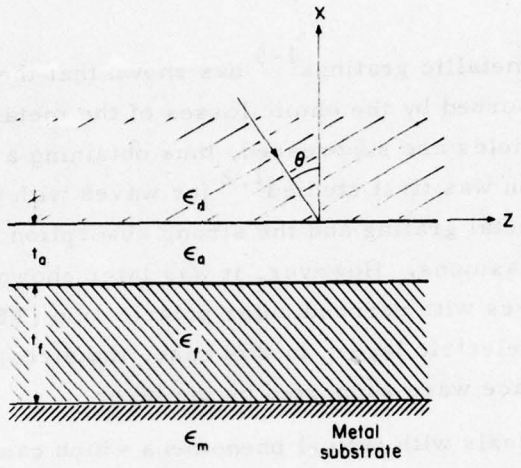


Fig. 1. Configuration of the layered structure.

as to represent lossy media. The former will then refer to a layer of thickness t_f having dielectric losses whereas the latter will represent a metal substrate having conduction losses.

For the situation shown and assuming a time variation $\exp(-i\omega t)$, the reflectivity at $z = 0$ for the incident wave can be expressed in the form

$$\rho(k_x) = \frac{(k_{zd} + iAk_{za})(k_{zf} - iFk_{zm})k_{za} - (k_{za} + iAk_{zd})(k_{zm} - iFk_{zf})k_{zf}}{(k_{zd} - iAk_{za})(k_{zf} - iFk_{zm})k_{za} + (k_{za} - iAk_{zd})(k_{zm} - iFk_{zf})k_{zf}} \quad , \tag{1}$$

where

$$A = A(k_x) = \tan k_{za} t_a \quad , \tag{2}$$

$$F = F(k_x) = \tan k_{zf} t_f \quad , \tag{3}$$

$$k_{zj} = [(2\pi/\lambda)^2 \epsilon_j - k_x^2]^{1/2} \quad , \tag{4}$$

for $j = d, a, f$ or m.

Here λ is the wavelength in vacuum, k_{zj} is the transverse wavenumber in the j-th medium, while k_x is the longitudinal wavenumber which is related to the incidence angle θ by

$$k_x = (2\pi/\lambda)\sqrt{\epsilon_d}\,\sin\theta \quad . \tag{5}$$

For a lossless configuration, the metal substrate implies that the imaginary part of ϵ_m becomes infinite while all other dielectric constants are finite and real. In that case, k_{zm} becomes infinite and Eq. (1) simplifies to

$$\rho(k_x) = \frac{(Ak_{zf} + Fk_{za})k_{zd} + i(AFk_{za} - k_{zf})k_{za}}{(Ak_{zf} + Fk_{za})k_{zd} - i(AFk_{za} - k_{zf})k_{za}} \quad . \tag{6}$$

While k_x is a real variable for incident plane waves, as in Eq. (5), the function $\rho(k_x)$ in Eq. (6) can be extended into the entire complex k_x plane. Thus, $\rho(k_x)$ has zeros and poles in this complex k_x plane, *which we denote by* k_{xo} *and* k_{xp}, *respectively.* The latter are associated with surface waves (k_{xp} real) and leaky waves (k_{xp} complex) that can be supported by the given configuration (see Reference 8).

In the present study, our concern is with the zeros of Eqs. (1) and (6); in particular we show that $\rho(k_x)$ in Eq. (1) may have a zero for real values of k_x and θ, i.e., a Brewster-type behavior occurs if the parameters are properly chosen. For this purpose, we recall that the configuration in Fig. 1 can support a leaky-wave having small attenuation along x if the dielectric constants satisfy $\epsilon_d > \epsilon_f > \epsilon_a$. This is the familiar situation of prism couplers,[8,9] except that now we have introduced a metal substrate to show the effect of conduction losses. In such cases, the significant leaky-wave pole occurs at

$$k_{xp} = \beta + i\alpha \quad , \tag{7}$$

where $\alpha \ll \beta$. A strong wave interaction is then known to occur if a beam is incident at an angle θ so as to satisfy the phase-match condition $\beta = k_x$, where k_x is given by Equation (5).

For lossless structures, the form of Eq. (6) suggests that, for values of k_x close to β, $\rho(k_x)$ can be approximated by[7]

$$\rho(k_x) = r(k_x)\frac{k_x - k_{xo}}{k_x - k_{xp}} \quad , \tag{8}$$

where $r(k_x)$ is slowly varying. In particular, for sufficiently small leakage, i.e., $\alpha \ll \beta$, as assumed here, $r(k_x) = R$ is constant for k_x close to β. Furthermore, for

lossless cases, it is obvious that the magnitude of $\rho(k_x)$ must be unity for plane-wave incidence because the metal substrate is perfectly reflecting and no absorption is assumed to occur anywhere. It is then easy to verify[7,9] that $|R| = 1$ and $k_{xo} = k_{xp}^*$, where the asterisk in k_{xp}^* denotes the complex conjugate value.

If small losses are introduced, the locations of both the pole (k_{xp}) and the zero (k_{xo}) will change, so that we may write Eq. (8) in the form

$$|\rho(k_x)| = \left| \frac{k_x - k_{xo}(\delta)}{k_x - k_{xp}(\delta)} \right| \quad , \tag{9}$$

where δ denotes a small-loss variable, which will be defined later. However, it is important to note that $k_{xo}(0) = k_{xp}^*(0)$, in agreement with the previous discussion. Also, by using calculations based on the exact expression, Eq. (1), we have found that the approximations involved in Eqs. (8) and (9) are extremely accurate for the small losses that occur in actual practice. This is in agreement with previous results obtained in acoustics,[7] as well as with a similar approximation that was used in the context of metallic gratings.[3-5]

Consider now the variation in the complex plane of both $k_{xo}(\delta)$ and $k_{xp}(\delta)$ as a function of δ, as shown for a typical case in Figure 2. In this case, it was assumed that the substrate is perfectly conducting and losses occur only in the layer of thickness t_f, which now has a complex dielectric constant given by

$$\varepsilon_f = \varepsilon_f'(1 + i \tan \delta_f) \quad . \tag{10}$$

Here ε_f' is real and $\tan \delta_f$ is the dielectric loss factor. The variation of $k_{xo}(\delta_f)$ and $k_{xp}(\delta_f)$ can be easily found by using a perturbation solution of Eq. (1), which was verified to be very accurate by means of computer calculations. However, these details are omitted here. In the lossless limit ($\tan \delta_f = 0$), the pole and zero are in conjugate locations, but they both move upward in the complex k_x plane as δ_f increases. In particular, at some critical value δ_{cf} of δ_f, the locus of $k_{xo}(\delta_f)$ intersects the real k_x axis at β_{cf}. This implies that an incidence angle θ_{cf} exists which satisfies Eq. (5), so that

$$\beta_{cf} = (2\pi/\lambda)\sqrt{\varepsilon_d} \sin \theta_{cf} \quad . \tag{11}$$

It therefore follows that $\rho(k_x) = \rho(\beta_{cf})$ vanishes if $\delta_f = \delta_{cf}$ and $\theta = \theta_{cf}$, i.e., reflection is entirely suppressed.

Because of the leaky-wave nature of the situation, the total absorption phenomenon can be attributed entirely to the strong interaction that occurs when the incident wave is phase-matched to the leakage angle. This is therefore the same situation that may

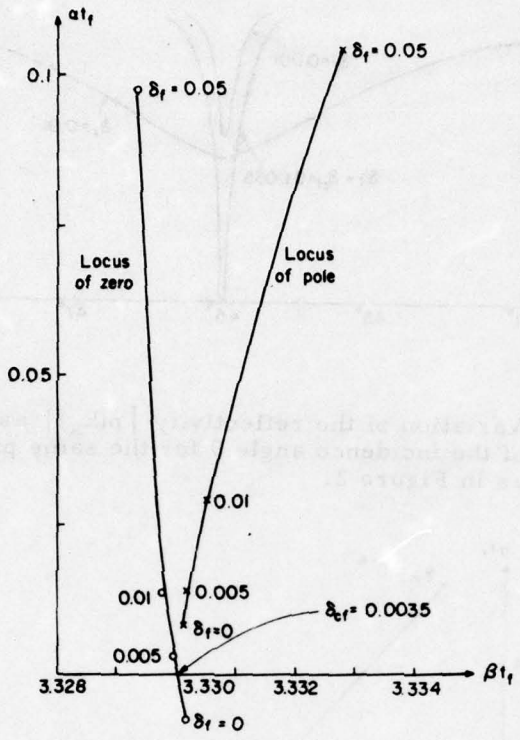


Fig. 2. Location of the leaky-wave pole and its associated zero in the complex k_x plane, as a function of δ_f, for $\epsilon_d = 4$, $\epsilon_a = 1$, $\epsilon_f = 3(1 + i \tan \delta_f)$, $t_a = 0.33\lambda$ and $t_f = 0.375\lambda$. The metal substrate is assumed to be a perfect conductor.

occur at the interface between a liquid and a solid in acoustics,[7] or in the prism-coupler configuration in optics.[9] As the angle θ is scanned about θ_{cf}, the reflectivity varies as shown in Fig. 3, which is quite similar to Fig. 13 of Reference 7. It is noted that the half-power points for the reflected field determine a very narrow range of strong absorption, so that $|\rho(k_x)|$ recovers quickly to unity not far from θ_{cf}.

If losses occur in the metal substrate rather than in the dielectric, we take

$$\epsilon_m = 1 + i \cot \delta_m \quad , \tag{12}$$

where now $\tan \delta_m = \omega \epsilon_o / \sigma$ represents the conduction loss factor. By using the same analytic approach as in the case of dielectric losses, we find that the loci of the pole and zero are now given by Figure 4. Again, the locus of k_{xo} intersects the real k_x axis at β_{cm} for a critical value δ_{cm} of δ_m. As in the case of dielectric losses, Fig. 5 now

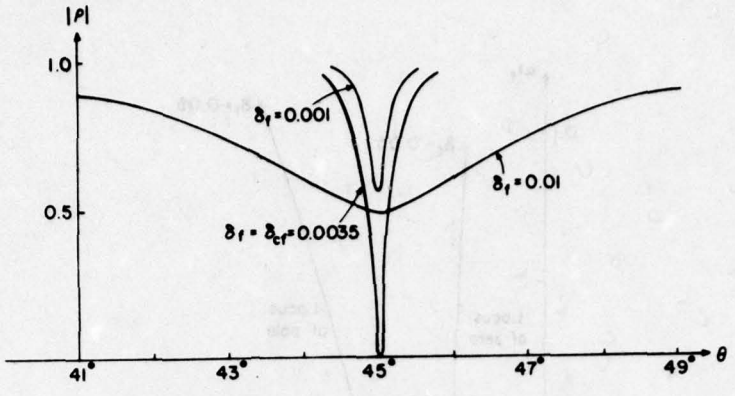
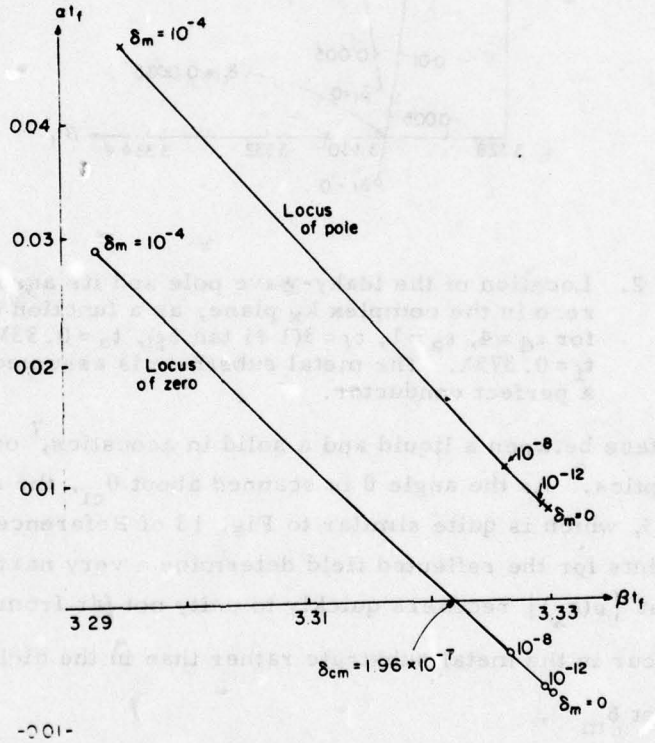


Fig. 3. Variation of the reflectivity $|\rho(k_x)|$ as a function of the incidence angle θ for the same parameters as in Figure 2.



-001-

Fig. 4. Location of the leaky-wave pole and its associated zero in the complex k_x plane, as a function of δ_m. The parameters are the same as in Fig. 2, except that now $\varepsilon_f = 3$ and the metal substrate is lossy, with $\varepsilon_m = 1 + i \cot \delta_m$.

shows that the variation of the reflectivity $\rho(k_x)$ around the corresponding angle θ_{cm} is such that $\rho(k_x)$ vanishes at θ_{cm} but recovers to unity not far from that critical angle.

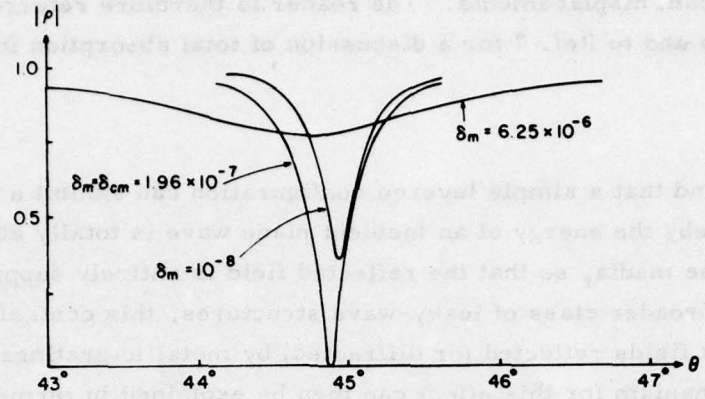


Fig. 5. Variation of the reflectivity $|\rho(k_x)|$ as a function of the incidence angle θ for the same parameters as in Figure 4.

C. Discussion

The behavior displayed in Figs. 2 to 5 is typical of the situation encountered when a plane wave is incident on a leaky-wave structure satisfying the small leakage condition $\alpha << \beta$. Thus, this behavior holds for TM modes as well as for the TE modes considered here; by extension, $\rho(k_x)$ can vanish for parameters ε_d, ε_a, ε_f, t_a and t_f that may be different from those chosen here. Also, more than four media may be involved and losses may occur in more than one of them. However, in all cases, the complete suppression of the reflected wave would occur only if both conditions $\delta = \delta_c$ and $\theta = \theta_c$ are satisfied simultaneously.

On the other hand, the leaky-wave configuration need not be restricted to layered media. In fact, the metallic gratings already studied in this context[1-5] are but particular cases of leaky-wave structures. Specifically, the conduction losses of metallic gratings are needed to provide the variation discussed after Eq. (12), while the presence of grooves and their periodicity is needed to provide a diffraction order to which an incident wave can be phase-matched. Phrased differently, these metallic gratings are structures that support waves of the leaky variety. As such, their total absorption properties follow the same criteria that have been presented here, provided only that they can support a leaky wave that satisfies the slow leakage condition $\alpha << \beta$.

Of course, the mathematical analysis of periodic structures, in general, and of metallic gratings, in particular, is considerably more complicated than the exposition given here. However, the basic leaky-wave feature, which is common to both layered

media and periodic structures, has already been studied and demonstrated by Hessel and Oliner[10] in the context of Wood anomalies, and by Tamir and Bertoni[9] in the context of bounded-beam displacements. The reader is therefore referred to their work for further details and to Ref. 7 for a discussion of total absorption in the analogous acoustic case.

D. Conclusion

We have found that a simple layered configuration can exhibit a Brewster-type phenomenon whereby the energy of an incident plane wave is totally absorbed by the ohmic losses in the media, so that the reflected field is entirely suppressed. Viewed as a model for a broader class of leaky-wave structures, this configuration explains the suppression of fields reflected (or diffracted) by metallic gratings with conduction losses. The mechanism for this effect can then be explained in terms of the strong interaction that occurs when the incident wave is phase-matched to the leaky-wave field that may be supported by either layered media or by periodic structures.

Consistent with the leaky-wave interpretation presented here, we find that total absorption of the incident energy may be due to either metallic or dielectric losses and it may occur in both TE and TM modes. Furthermore, we note that the leaky-wave behavior discussed in electromagnetic terms here covers not only effects that have been observed along optical gratings, but is analogous to phenomena that have been observed and studied in acoustics.

National Science Foundation
ENG74-23908

V. Shah and T. Tamir

REFERENCES

1. M.C. Hutley and V.M. Bird, Opt. Acta, Vol. 20, p. 771 (1973).
2. D. Maystre and R. Petit, Opt. Commun., Vol. 17, p. 196 (1976).
3. D. Maystre and M. Nevière, Nouv. Rev. Opt., to be published.
4. M. Nevière, D. Maystre and P. Vincent, Nouv. Rev. Opt., to be published.
5. M.C. Hutley and D. Maystre, Opt. Commun., to be published.
6. F.L. Becker and R.L. Richardson, J. Acoust. Soc. Am., Vol. 51, p. 1609 (1971).
7. H.L. Bertoni and T. Tamir, Appl. Phys., Vol. 2, p. 157 (1973).
8. T. Tamir, Nouv. Rev. Opt., Vol. 6, p. 273 (1975).
9. T. Tamir and H.L. Bertoni, Opt. Soc. Am., Vol. 61, p. 1397 (1971).
10. A. Hessel and A.A. Oliner, Appl. Opt., Vol. 4, p. 1275 (1965).

SHEAR WAVE PROPAGATION IN SYMMETRIC HEXOGONAL HONEYCOMB CORES

H. L. Bertoni and S. K. Park

Honeycomb panels, consisting of a hexagonal core structure bonded to metal face plates, are strong lightweight construction elements. These elements are used in fabricating containers for housing electronic equipment, in airplane construction, and other applications.

The core consists of an array of thin-walled, hollow cylinders of hexagonal cross-section, as shown in Fig. 1, and resembles a honeycomb. The core material consists of waxed cardboard in the case of equipment containers, while in airplane construction it is usually metal. Face plates are attached to both ends of the core using epoxy or other adhesive bonding materials (one of the face plates is shown in Figure 1).

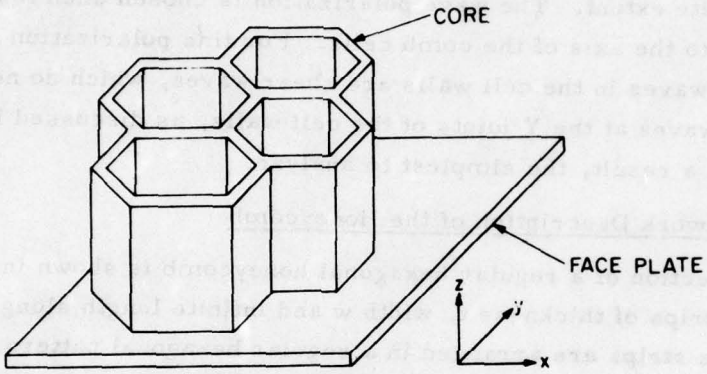


Fig. 1. Honeycomb core attached to one face plate.

During fabrication and use, defects can occur within the panel, and are therefore not detectable by visual inspection. One type of defect is a region of a face plate that is not bonded to the core. Another defect is the presence of water in some of the cells, which can degrade a cardboard core and weaken the epoxy or adhesive bond between core and face plate. Other defects are buckled or fractured cell walls.

Currently, acoustic non-destructive evaluation (NDE) techniques are used to locate the defects cited above. An acoustic beam in a fluid is directed at normal incidence on the panel and received at the opposite face. The beam is then scanned in a raster fashion over the area of the panel. Unbounded regions interrupt the transmission while water in cells increases the transmission. Since the size of the acoustic beam is on the order of the cell diameter, the raster scan inspection procedure is very lengthy for a large panel. A great reduction in inspection time could be achieved by propagating the acoustic waves parallel to the face plates, rather than perpendicular to them. In this way, one would inspect an entire strip along the panel at one time. Reinspecting installed panels with only one face exposed would require propagation

along the panel.

Low frequency studies of propagation along honeycomb panels[1] have been carried out with the aim of NDE inspection in mind. However, it was recognized that successful NDE inspection would require the use of high frequencies for which the acoustic wavelengths would be on the order of the cell diameter. Establishment of a sensitive test procedure requires a knowledge of the propagation characteristics of the modes guided by honeycomb panels. From such knowledge, one can determine the frequency and mode polarization most useful for inspection. These choices must be made in relation to the geometric design and elastic properties of materials used in the panel.

As a first step in the analysis of the modes guided by honeycomb panels, we have computed the properties of waves propagating transverse to the cell axis in a honeycomb of infinite extent. The wave polarization is chosen such that the particle motion is parallel to the axis of the comb cells. For this polarization and direction of propagation the waves in the cell walls are shear waves, which do not excite the Lamb or flexural waves at the Y joints of the cell walls, as discussed below. This polarization is, as a result, the simplest to analyze.

A. Equivalent Network Description of the Honeycomb

The cross-section of a regular hexagonal honeycomb is shown in Figure 2. The core consists of strips of thickness t, width w and infinite length along z (out of plane of the paper). The strips are arranged in a regular hexagonal pattern. For realistic honeycombs, $t << w$, so that the strips are thin compared to wavelength, even for frequencies where the periodic nature of the core is significant. With the assumption that t is small compared to wavelength, the elastic fields in the strips are those associated with the three lowest modes of a plate, namely the Lamb mode (L), S-H plate mode (S) and the flexual mode (F). In addition, near the joints between the strips the next higher flexual mode F_1, which is evanescent away from the junction, will be significant.[2] For t small, the other evanescent plate modes will not be strongly excited. Finally, thin plate or plane strain approximations can be used for the modes.[3]

A unit cell of the periodic core is indicated by the dashed curve in Figure 2. The entire periodic core can be generated by translations of the unit cell along the basis vectors $\underline{a}_1$ and $\underline{a}_2$, shown in Fig. 2, of the periodic structure. The cell is seen to be composed of plate segments coupled at two Y joints. A primary part of the study of core waves therefore involves the coupling of the plate waves at the Y joint.

In order to satisfy the boundary conditions at the Y joint, it is necessary that all plate waves have the same phase dependence $\exp(-jk_z z)$ along z. The variation of the

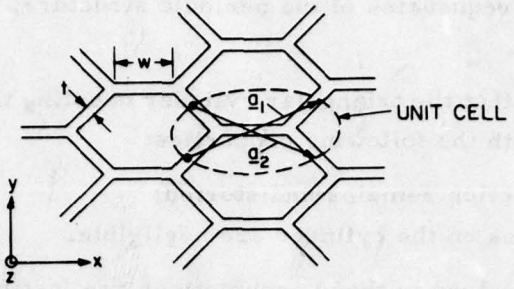


Fig. 2. Cross-section of regular hexagonal honeycomb core. Unit cell is indicated by dashed ellipse.

plate waves transverse to z can be represented by equivalent transmission lines.[5-9] For propagation in the x-y plane, $k_z = 0$. In this case an S-H mode the plate, whose particle motion is along z, will not couple to the Lamb or flexural modes at the Y joint, since their particle motion is perpendicular to z.

For the lowest S-H mode of a plate, the particle velocity and stress have no variation through the thickness of the plate. For propagation perpendicular to z, the particle velocity has only the component v_z along z, whereas the force or stress across a surface perpendicular to the direction of propagation has only the component T_{tz} along z. Propagation of this S-H mode can be modeled by a transmission line whose voltage V and current I represent the particle velocity v_z and stress T_{tz} via the relations.[8]

$$\left.\begin{aligned} v_z &= \frac{1}{\sqrt{t\mu\rho}}\, V \\ T_{tz} &= -\sqrt{\frac{\mu\rho}{t}}\, I \end{aligned}\right\} \quad . \tag{1}$$

Here ρ is the mass density and μ is the shear modulus. The impedance of the transmission time is

$$z_s = \sqrt{\rho\mu} \tag{2}$$

and its propagation wavenumber is

$$k_s = \omega\sqrt{\frac{\rho}{\mu}} \quad . \tag{3}$$

In solving for the coupling properties of the Y junction, all of the evanescent S-H waves are neglected. This approximation is consistent with the thin plate, or plane strain, approximation. By neglecting these higher modes we introduce a small phase error in the waves excited at the junction. This phase error will cause a slight

shift in the stop-band frequencies of the periodic structure, which for NDE applications is not significant.

We next assume that the triangular cylinder coupling the plates at the Y junction (see Fig. 3) behaves with the following properties:

(1) the cross-section remains undistorted;

(2) inertial forces on the cylinder are negligible.

The approximations involved in these assumptions are justified by the fact that the width of the sides of the triangular cylinder are equal to the plate thickness t. In the thin plate approximation, the plate thickness does not change so that the triangle cannot distort. Also, the inertia of the mass in the triangular cylinder is less than one half that in a plate segment of width equal to the plate thickness.

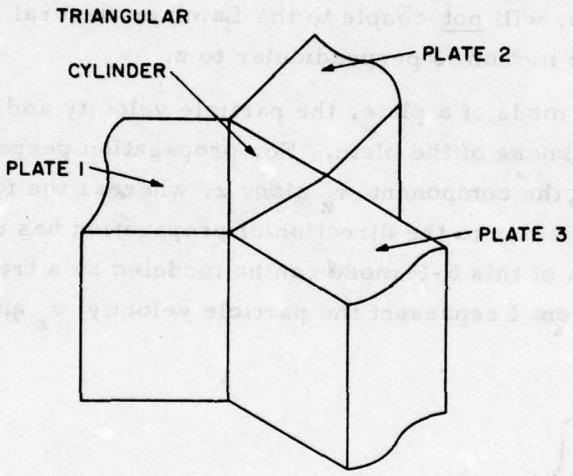


Fig. 3. Triangular cylinder coupling three plates.

With the foregoing assumptions, the dynamic conditions at the faces of the triangular cylinder are:

(1) the forces on the three sides of the triangle must sum to zero;

(2) the particle velocities at the center of each side of the triangle must be equal.

Since the force T_{tz} is measured by current, and the particle velocity v_z by voltage,[8,9] condition (1) above requires that the sum of the currents be zero, while condition (2) requires that the voltages be equal. The Y junction is then represented by the star connection of transmission lines shown in Figure 4.

Knowing the network representation of each Y junction, the equivalent network for a unit cell can be obtained. This network is drawn in Fig. 5 and is seen to consist

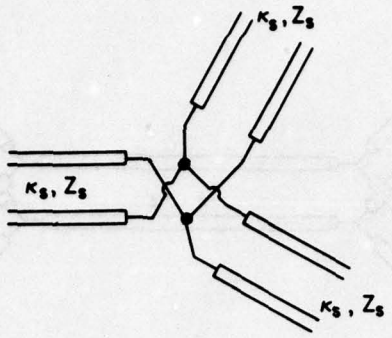


Fig. 4. Network for Y junction of three plates representing S modes coupling at normal incidence ($k_z = 0$).

of four ports at which the voltages and currents V_i, I_i (i=1-4) are defined. The four currents can be related via an impedance matrix. Because of the symmetry of the network, this relationship takes the form

$$\begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} = \begin{bmatrix} Z_{11} & Z_{12} & Z_{13} & Z_{13} \\ Z_{12} & Z_{11} & Z_{13} & Z_{13} \\ Z_{13} & Z_{13} & Z_{11} & Z_{12} \\ Z_{13} & Z_{13} & Z_{12} & Z_{11} \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \\ I_3 \\ I_4 \end{bmatrix} , \tag{4}$$

where Z_{11}, Z_{12} and Z_{13} are given by

$$Z_{11} = jZ_s \frac{1-10\tan^2\theta+7\tan^4\theta}{6\tan\theta(2\tan^2\theta-1)} \tag{5}$$

$$Z_{12} = jZ_s \frac{1-5\tan^2\theta}{6\cos^2\theta\tan\theta(2\tan^2\theta-1)} \tag{6}$$

and

$$Z_{13} = jZ_s \frac{1}{6\cos^4\theta\tan\theta(2\tan^2\theta-1)} \tag{7}$$

with

$$\theta = k_s w/2 \quad . \tag{8}$$

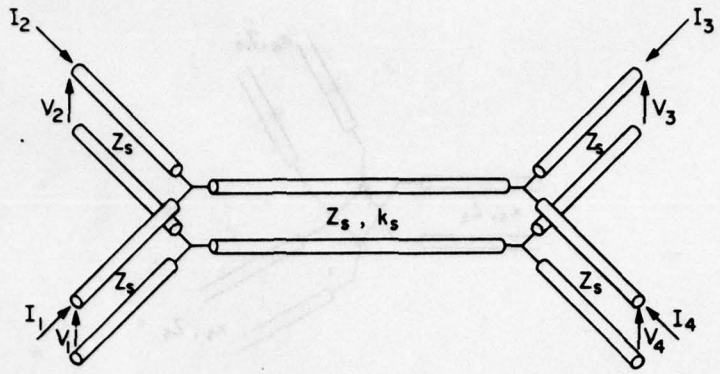


Fig. 5. Equivalent circuit representing a unit cell of the honeycomb.

B. Wave Propagation in the Periodic Honeycomb

We now consider the terminal voltages and currents V_i and I_i to be associated with a wave traveling through the core. Let

$$\underline{k} = \underline{x}_o k_x + \underline{y}_o k_y \tag{9}$$

be the wave vector of this wave. Also, let $\underline{a}_1$ and $\underline{a}_2$ be the basic translation vectors of the periodic structure, as shown in Fig. 2, and given by

$$\underline{a}_{1,2} = \frac{3}{2} w \ \underline{x}_o \pm \frac{\sqrt{3}}{2} w \underline{y}_o \tag{10}$$

The Floquet assumption for periodic structures then implies the following relation between the terminal voltages and currents

$$\left.\begin{aligned} V_3 &= V_1 e^{-j\underline{k}\cdot\underline{a}_1} \\ I_3 &= -I_1 e^{-j\underline{k}\cdot\underline{a}_1} \\ V_4 &= V_2 e^{-j\underline{k}\cdot\underline{a}_2} \\ I_4 &= -I_2 e^{-j\underline{k}\cdot\underline{a}_2} \end{aligned}\right\} \tag{11}$$

The minus sign before I_1 and I_2 is due to the fact the currents are assumed positive into the unit cell.

Substituting Eq. (11) into Eq. (4) one obtaines four linear homogeneous equations in terms of the four unknowns V_1, V_2, I_1 and I_2. Defining

$$\left.\begin{aligned}\psi_1 &= e^{-j\underline{k}\cdot\underline{a}_1}\\ \psi_2 &= e^{-j\underline{k}\cdot\underline{a}_2}\end{aligned}\right\} \tag{12}$$

the four equations may be written as

$$0 = \begin{bmatrix} Z_{11}-\psi_1 Z_{13} & Z_{12}-\psi_2 Z_{13} & -1 & 0\\ Z_{12}-\psi_1 Z_{13} & Z_{11}-\psi_2 Z_{13} & 0 & -1\\ Z_{13}-\psi_1 Z_{11} & Z_{13}-\psi_2 Z_{12} & -\psi_1 & 0\\ Z_{13}-\psi_1 Z_{12} & Z_{13}-\psi_2 Z_{11} & 0 & -\psi_2 \end{bmatrix} \begin{bmatrix} I_1\\ I_2\\ V_1\\ V_2 \end{bmatrix} . \tag{13}$$

In order for Eq. (13) to have a non-trivial solution, the determinant of the 4x4 matrix must vanish. Setting the determinant equal to zero gives the dispersion equation of the form

$$(Z_{13}^2 - Z_{12}^2)\cos^2\delta + 2Z_{13}(Z_{12}-Z_{11})\cos\delta\cos\gamma + (Z_{11}^2 - Z_{13}^2) = 0 \tag{14}$$

where

$$\left.\begin{aligned}\delta &= \frac{\sqrt{3}}{2} wk_y\\ \gamma &= \frac{3}{2} wk_x\end{aligned}\right\} \tag{15}$$

This equation may then be solved numerically for its root k_x as a function of k_y and ω or k_y as a function of k_x and ω . The root corresponds to a wave in the core whose polarization is found from the eigen vector associated with the root.

Because of the periodicity of the honeycomb, Eq. (13) will have multiple solutions. Let $\underline{u}_1$ and $\underline{u}_2$ be reciprocal lattice vectors

$$\underline{u}_{1,2} = \frac{2\pi}{3w}\,\underline{x}_o \pm \frac{2\pi}{\sqrt{3}w}\,\underline{y}_o \tag{16}$$

so that

$$\underline{u}_i \cdot \underline{a}_j = 2\pi\delta_{ij} \tag{17}$$

where δ_{ij} is the Kronecker delta. Thus, if a particular wavevector $\underline{k}$ satisfies Eq. (14), then for any integers m and n the wavevector $\underline{k} + m\,\underline{u}_1 + n\,\underline{u}_2$ will also

satisfy Equation(13). Also, due to the hexagonal symmetry of the honeycomb, a plot of k_x versus k_y for fixed ω will have six-fold symmetry about the origin.

Numerical computations have been carried out for the dispersion curves k_x versus k_y for various fixed values of ω . Because the S-H wave velocity is independent of the wall thickness t, and because the impedances of all transmission lines in Fig. 5 are the same, the results are independent of t. In addition, if the normalized frequency variable θ in Eq. (8) is used, and the wave numbers k_x and k_y are multiplied by $w/2$, the computations will apply to all materials and honeycombs.

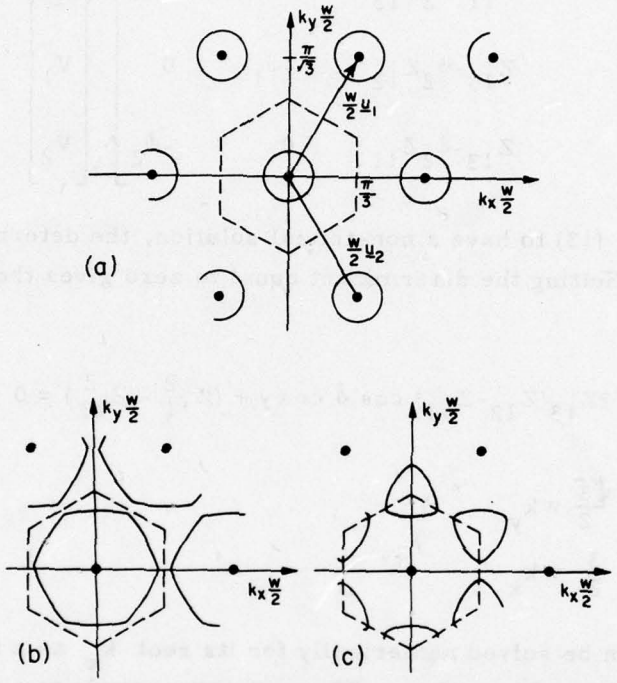


Fig. 6. Reciprocal lattice points and slowness curves for: a) frequency parameter $\theta << 0.6155$; b) θ just below 0.6155; and c) θ in the range $0.6155 < \theta < \pi/4 = 0.7854$.

Due to the symmetry and periodicity of the honeycomb, the first quadrant in the k_x-k_y plane of the first Brillonin zone is more than adequate to present the numerical results for the dispersion curves of k_x versus k_y. However, in order to explain the variation of the dispersion curves with frequency we have sketched the curves for several Brillonin zones in Figures 6(a), (b), (c). In Fig. 6(a) we have indicated several reciprocal lattice points by dots and the vectors $\underline{u}_1$ and $\underline{u}_2$. The dashed hexagon represents the first Brillonin zone. For low frequencies ($\theta << 0.6155$), the dispersion curve is very nearly a circle of radius $|\underline{k}| \frac{w}{2} = \sqrt{2}\,\theta_s$. This value of $|\underline{k}|$

corresponds to a wave velocity for the Block wave that is $1/\sqrt{2}$ times shear wave velocity of the honeycomb material. Because of the periodicity, the near-circles centered about each reciprocal lattice point are all loci of solutions of the dispersion relation, Eq. (14), as discussed after Equation (17).

As the frequency variable θ increase to just below 0.6155, the near-circular dispersion curves expand and distort into nearly hexagonal shapes, as indicated in Figure 6(b). At $\theta = 0.6155$, the tips of the hexagons touch. For $0.6155 < \theta < \frac{\pi}{4} = 0.7854$, the dispersion curves switch to being nearly triangular and centered about the apexes of the dashed hexagon. As θ approaches $\pi/4$, the dispersion curves shrink to the points at the apexes of the hexagon. For θ increasing above $\pi/4$, the change in the dispersion curves is reversed from that described above until for θ approaching $\pi/2$, the curves become nearly circular as in Figure 6(a). For θ increasing above $\pi/2$, the variation of the dispersion curves described above is repeated periodically with period $\pi/2$.

The actual dispersion curves in the first quadrant of the first Brillouin zone are plotted to scale in Fig. 7, and are seen to be in agreement with the discussion of Figure 6. One of the features of these curves that may important for NDE is the fact that in certain frequency ranges the waves are cut-off along $\pm x$, and the four other directions making angles of 60° to the $\pm x$ axis. These cut-off directions are ones parallel to a set of cell walls. *However, the waves are never cut-off along y*, and the six-fold related axes.

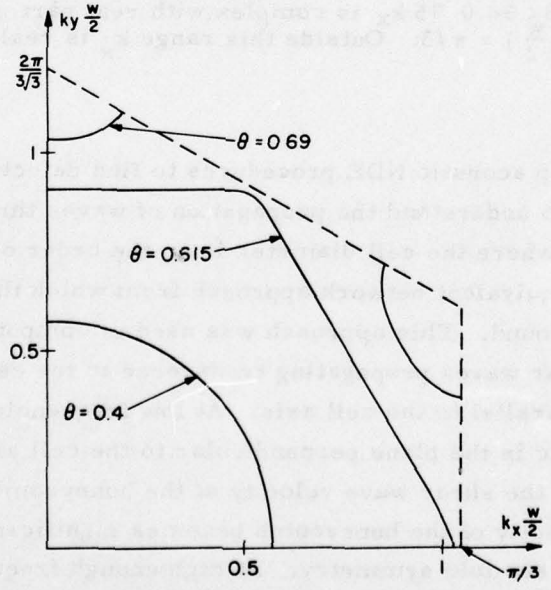


Fig. 7. Slowness curves in the first quadrant of the first Brillouin zone for several values of the frequency parameter θ.

In order to show the cut-off behavior for propagation along x, we have plotted k_x as a function of ω for the case $k_y=0$. This plot is shown in Fig. 8, where k_x is real for $0 < \theta < 0.6155$ and $.9553 > \theta > \pi/2$. In the range $0.6155 < \theta < 0.9553$, k_x is complex with real part given by $\mathrm{Re}(k_x \frac{w}{2}) = \pi/3$. The curves are symmetric about $\theta = \pi/4$. For θ outside the range plotted, the curves repeat with period $\pi/2$.

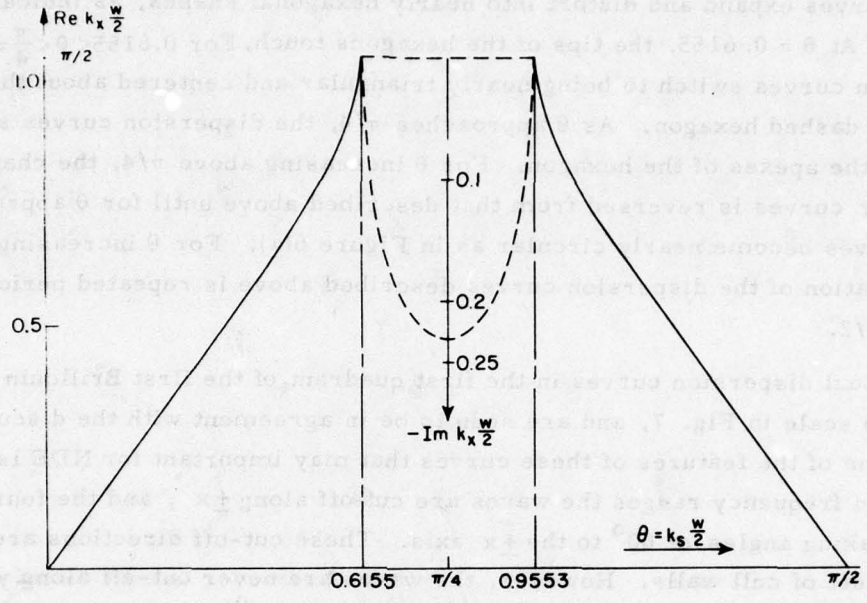


Fig. 8. Dispersion curve for propagation along x($k_y = 0$). In the range $0.6155 < \theta < 0.95$ k_x is complex with real part given by $\mathrm{Re}\,(k_x \frac{w}{2}) = \pi/3$. Outside this range k_x is real.

C. Summary

In order to develop acoustic NDE procedures to find defects in honeycomb panels it is necessary to understand the propagation of waves through the honeycomb, especially in the range where the cell diameter is on the order of the wavelength. We have developed an equivalent network approach from which the propagation characteristics can be found. This approach was used to compute the propagation characteristics for shear waves propagating transverse to the cell axis and polarized with particle velocity parallel to the cell axis. At low frequencies the wave propagation was nearly isotropic in the plane perpendicular to the cell axis with phase velocity $\frac{1}{\sqrt{2}}$ times lower than the shear wave velocity of the honeycomb material. At higher frequencies, the periodicity of the honeycomb becomes significant and the propagation is an isotropic with six-fold symmetry. At high enough frequencies, no propagation takes place in directions parallel to each of the sets of cell walls. However, in

the orthogonal directions propagation takes place for all frequencies.

Air Force Systems Command
FY76207702615

H. L. Bertoni and S. K. Park

National Science Foundation
ENG 75-00569

REFERENCES

1. R. B. Thompson, et al., "Dispersion of Flexural Elastic Waves in Honeycomb Sandwich Panels," JASA, Vol. 57, pp. 119-1127 (1975).

2. L. Cremer, M. Heckel and E. E. Ungar, "Structure-Borne Sound," (New York: Springer-Verlag, 1973) pp. 306-335.

3a. R. D. Mindlin, "Influence of Rotary Inertia and Shear on Flexural Motions of Isotropic Elastic Plates," J. Appl. Mech., pp. 31-38 (1951).

3b. T. R. Kane and R. D. Mindlin, "High Frequency Extensional Vibrations of Plates," J. Appl. Mech., pp. 277-283 (1956).

4. R. B. Thompson, et al., "Acoustic Delays of the Order of a Millisecond per Inch in Cellular Materials," IEEE Trans., Vol. SU-18, p. 184 (1971).

5. A. A. Oliner, H. L. Bertoni and R. C. M. Li, "A Microwave Network Formalism for Acoustic Waves in Isotropic Media," Proc. IEEE, Vol. 60, pp. 1503-1512 (1972).

6. A. A. Oliner, R. C. M. Li and H. L. Bertoni, "Catalog of Acoustic Equivalent Networks for Planar Interfaces," Proc. IEEE, Vol. 60, pp. 1513-1518 (1972).

7. R. C. M. Li, A. A. Oliner, H. L. Bertoni, "Microwave Network Analysis of Surface Acoustic Waveguides, I: Flat Overlay Guides," IEEE Trans. Sonics Ultrason., Vol. SU-24, No. 2, pp. 66-78 (March 1977).

8. S. Markman, R. C. M. Li, A. A. Oliner, H. L. Bertoni, "Microwave Network Analysis of Surface Acoustic Waveguides, II: Rectangular Ridge Guides," IEEE Trans. Sonics Ultrason., Vol. SU-24, No. 2, pp. 79-87 (March 1977).

9. S. Markman, "A Microwave Network Analysis of Ridge-Type Surface Acoustic Waveguides," Ph. D. Dissertation, Polytechn. Inst. of New York (1973).

OBSERVATION OF THE LEAKY-WAVE FIELD IN AN ACOUSTICALLY-EXCITED LAYERED STRUCTURE

D. A. Davids and J. O'Brien

A. Introduction

The existence of a leaky-wave has been used to explain the observed lateral displacement of an acoustic beam reflected from a solid substrate,[1] the operation of acoustic wedge transducers,[2] and as a design principle in a prism device for nondestructive evaluation.[3] Although the beam displacement and the concomitant null region in the reflected beam have been observed and documented,[4] the existence and predicted functional dependence of the leaky-wave has not been given an explicit experimental verification.

In order to provide a clearer test of the predicted properties of the leaky-wave, an experiment was designed in which a wedge transducer is used to launch on the surface of a polished aluminum substrate a Rayleigh wave which is then made to impinge on a layered structure consisting of a glass slab of dense flint separated from the substrate by a fluid film of controlled thickness. Refer to Figure 1. The upper surface

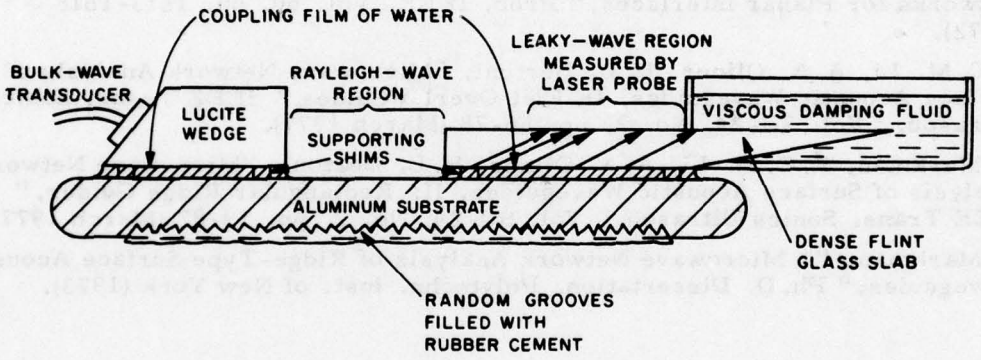


Fig. 1. System used to observe leaky-wave characteristics by means of a heterodyne-type laser probe.

of the glass slab is of optical quality and is aluminized to provide a mirror surface. The displacement of this upper surface at the acoustic excitation frequency is then measured by means of a heterodyne laser probe the operation of which is described by Whitman and Korpel.[5] The complete structure is located on an x-y positioner which enables a determination of the acoustic excitation as a function of either lateral or transverse coordinates. The thickness of the fluid-film is established by plastic shims that are located well outside of the region of the acoustic beam. The slab is terminated in a reservoir filled with a viscous damping fluid which approximately "index matches" the acoustic wave in the glass, the function of which is to absorb energy from the

acoustic wave and thus reduce reflections and the consequent standing waves. Unwanted surface-wave reflections on the aluminum substrate were reduced to a negligible point by rounding the edges of the substrate and engraving the underside with a random pattern of grooves that were filled with rubber cement.

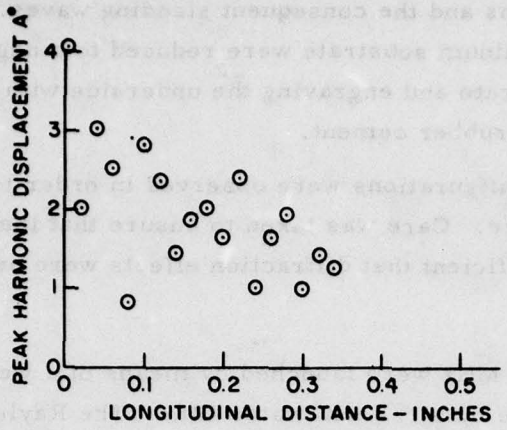
A number of transducer wedge configurations were observed in order to obtain a Rayleigh beam of nearly Gaussian profile. Care was taken to ensure that the widths of wedges, substrates, and slabs were sufficient that diffraction effects were negligible.

B. Procedure

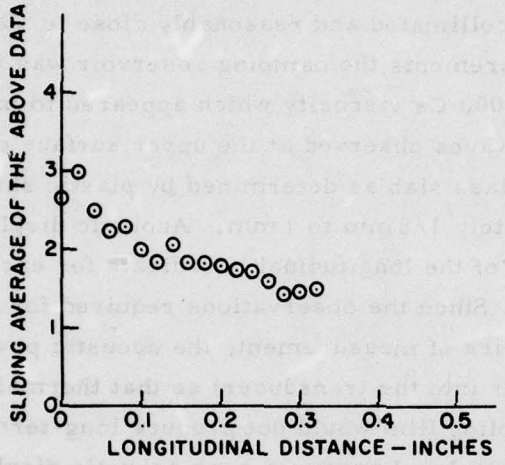
Continuous Rayleigh waves of f = 1 MHz were launched by means of a wedge transducer coupled to the aluminum substrate by a 1/2 mm water film. The Rayleigh beam was observed and confirmed to be well collimated and reasonably close to Gaussian in profile. During all experimental measurements the damping reservoir was filled with Dow Corning type 200 silicone fluid of 1000 Cs viscosity which appeared to most dramatically reduce the standing acoustic waves observed at the upper surface of the glass slab. Water film thickness under the glass slab as determined by plastic shim stock was varied over the range of approximately 1/8 mm to 1 mm. Acoustic displacement of f = 1 MHz was measured as a function of the longitudinal coordinate for each of the reported values of fluid film thickness. Since the observations required for a single fluid film thickness involved several hours of measurement, the acoustic power levels were kept low (< 100 mw electrical power into the transducer) so that thermal effects in the PZT transducer and its fluid coupling film would not produce long-term effects in the acoustic excitation. This excitation level produced peak acoustic displacements at the upper surface of the glass slab in the range of 0.1 to 5 Å which, unfortunately, restricted the range over which the decay could be observed. During each experimental run previously measured points were constantly rechecked to ensure long-term repeatability. Where standing waves were insignificant, data points were plotted on semi-log paper and exponential decay constants determined. Where standing waves were present, data points were averaged over one period and the "sliding-average" value plotted semi-logarithmically. This sliding average produced considerable data smoothing of the 4-5 data points per acoustic wavelength, while introducing little error since in the cases where this technique was used, the decay length was large compared with an acoustic wavelength. Figure 2 illustrates cases where standing waves are either absent or present, and in the latter case the effect of the sliding average on the measured data.

C. Results

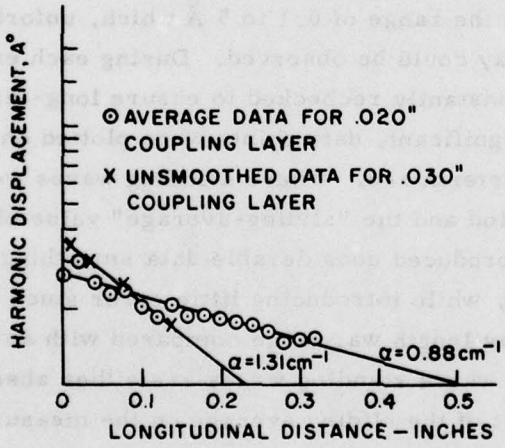
The results of the measurements are illustrated in Fig. 3, where the predictions of Hou and Bertoni appear as solid lines. Their calculations were for a steel rather than an aluminum substrate but this should make little difference in the result. We



A. Measured harmonic acoustic displacement in the presence of standing waves (worst case) for .020" coupling film thickness.



B. Sliding average over one acoustic wavelength. The phase constant β of the leaky wave used in averaging was measured to be $\beta = 3.55\,cm^{-1}$.



C. Semilog plot of averaged displacement for above case denoted by ⊙ The line corresponds to a decay constant of $\alpha = 0.88\,cm^{-1}$. Also indicated is unsmoothed data for .030" coupling layer. The line represents a decay constant of $\alpha = 1.31\,cm^{-1}$.

Fig. 2

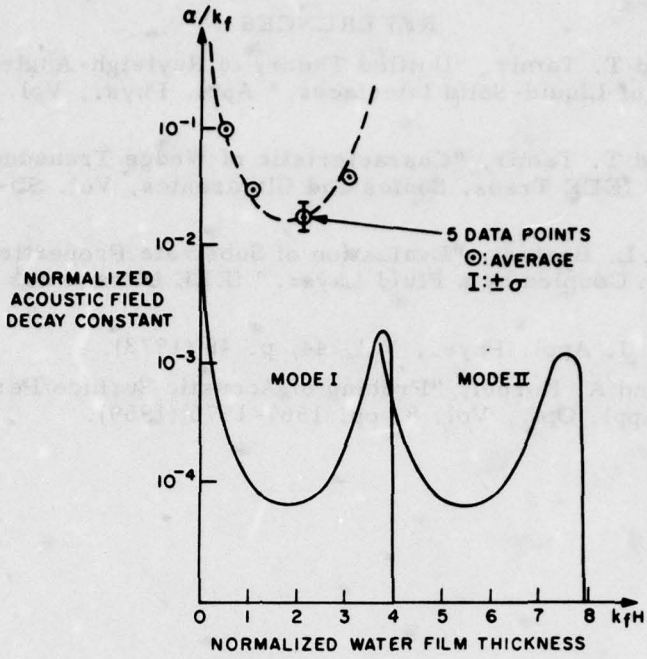


Fig. 3. Comparison of experimental measurements for water layer between aluminum and dense flint glass compared with the predictions of Hou and Bertoni[3] for a water layer between steel and dense flint glass (solid line). The dotted curve represents a scaled version of the predicted result (scale factor x 200). Both decay constant α and water film thickness H are normalized to the acoustic wave number in water, k_f.

observe that our decay constants are two hundred times the decay constants predicted by the theory. For comparison we have multiplied the Hou and Bertoni predicted values by approximately 200 and plotted the result as a dotted line. The functional dependence on film thickness is then see seen to be quite good. The phase constant of the leaky-wave on the surface of the flint-glass slab was observed to be $\beta = 3.55\,cm^{-1}$.

D. Conclusions

Under the conditions detailed above, we have observed exponentially decaying leaky-wave fields with decay constants which differ by a large factor from the values numerically predicted by Hou and Bertoni, but, which when a propriately scaled, show a remarkable agreement with the predicted functional dependence on film thickness. At this time we have not as yet been able to identify the source of the scaling discrepancy.

Joint Services Technical Advisory Committee
F44620-74-C-0056

National Science Foundation
ENG76-80671

D. A. Davids and J. O'Brien

REFERENCES

1. H.L. Bertoni and T. Tamir, "Unified Theory of Rayleigh-Angle Phenomena for Acoustic Beams of Liquid-Solid Interfaces," Appl. Phys., Vol. 2, pp. 157-172 (1973).

2. H.L. Bertoni and T. Tamir, "Characteristic of Wedge Transducers for Acoustic Surface Waves," IEEE Trans. Sonics and Ultrasonics, Vol. SU-22, pp. 415-420 (1975).

3. Y.L. Hou and H.L. Bertoni, "Evaluation of Substrate Properties Using an Acoustic Beam in a Prism Coupled by a Fluid Layer," IEEE Ultrasonics Symposium Proc. (1975).

4. W.G. Neubauer, J. Appl. Phys., Vol. 44, p. 48 (1973).

5. R.L. Whitman and A. Korpel, "Probing of Acoustic Surface Perturbations by Coherent Light," Appl. Opt., Vol. 8, pp. 1567-1576 (1969).

LEAKY SAW AND GUIDED LIGHT BRAGG INTERACTION IN $LiNbO_3$

F. Freyre and W-C. Wang

The results of a theoretical analysis of Bragg interaction of optically guided light and leaky acoustic surface waves are presented. These results are compared to those of a similar analysis for YZ Rayleigh waves. For TM optical modes, X propagating leaky waves on metallized 64° rotated Y cut $LiNbO_3$ interact nearly twice as strongly at the surface as the free surface Rayleigh waves. However, X propagating leaky waves on free surface 41° rotated Y cut $LiNbO_3$ have near zero interaction for both TM and TE modes.

A. Introduction

Leaky waves on rotated Y cut $LiNbO_3$, were described by Yamanouchi and Shibayama.[1,7] They found that X traveling waves on 41° rotated Y cut free surface $LiNbO_3$ and 64° Y cut metallized $LiNbO_3$ have zero decay constants. Thus they are true surface waves with no bulk wave components. They also found that the waves have large electro-mechanical coupling coefficients. In our analysis of the Bragg acousto-optic interactions of optical surface waves with these waves (and of Rayleigh waves)[8] we shall assume the existence of a thin optical waveguide on the surface of the $LiNbO_3$. This guide will have been fabricated by metal ion indiffusion[3] or Li_2O outdiffusion[6] with a resulting tapered refractive index profile. In some cases we shall consider the surface to be thinly metallized. We shall also assume that the diffusion process shall not have significantly affected the elasto-optic and electro-optic coefficients of the $LiNbO_3$.

Figure 1 shows the relative orientation of the crystal, light and SAW.

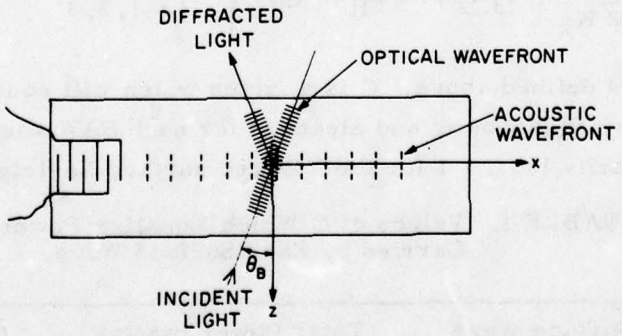


Fig. 1. Top surface of crystal showing relative directions of guided lightwaves and leaky surface waves.

B. Acoustic Analysis

For TE^2 and TM light incident at the Bragg angle, the Bragg diffraction efficiency, ζ, is approximately given by

$$\zeta = \sin^2 \frac{\Gamma_{nm} L}{\cos \theta_B} \tag{1}$$

where L is the acoustic beam width and Γ_{nm} is the coupling constant for mode m to mode n diffracted light. θ_B is the Bragg angle. The expression for the coupling constant is

$$\Gamma_{nm} = \frac{\pi n^3}{2\lambda_o} \frac{\int_o^{-\infty} [P_{IJ} S_J(y) + r_{Ij} E_j^A(y)]\ \mu_m(y)\, \mu_n(y)\, dy}{\int_o^{-\infty} \mu_n^2(y)\, dy} \qquad \begin{matrix} J = 1 \text{ to } 6 \\ j = 1 \text{ to } 3 \end{matrix} \tag{2}$$

I = 1 for the X traveling leaky waves and TE modes, 3 for the Z traveling Rayleigh waves and TE modes, and 2 for TM modes with the Rayleigh or leaky waves. $\mu_m(y)$, $\mu_n(y)$ are the optical electric field distributions for TE modes and the optical magnetic field distributions for TM modes. n is the refractive index. λ_o is the free space wavelength of the light. P_{IJ}, r_{Ij} represent elasto-optic and electro-optic tensor elements, respectively, in the crystal cut coordinate system. $S_j(y)$ is the jth strain component of the SAW. E_j^A is the jth electric field component of the SAW. For TM modes we have assumed E_y is much greater than E in the light propagation direction which is generally true.

We define the Overlap Factor (O.F.) as follows:

$$\text{O.F.} = \frac{C}{A} \left\{ \frac{\pi n^3}{2\sqrt{2} K_A} [P_{IJ} S_J(y) + r_{Ij} E_j^A(y)] \right\} \qquad \begin{matrix} J = 1 \text{ to } 6 \\ j = 1, 2, 3 \end{matrix}$$

Subscript I was defined above. C is a factor which will equalize the total power flow per unit beam width (acoustic and electric) for each SAW assuming equal wavenumber. We arbitrarily let C = 1 for the YZ free surface Rayleigh Wave (see Table I)

TABLE I. Values of C Which Equalize Power Carried by Each Surface Wave.

Surface Wave	Total Power (watts)	C
41° F.S. Leaky	$10.95 \times 10^{18} A^2 LK_A$	.0068
64° Met. Leaky	$8.69 \times 10^{15} A^2 LK_A$	.24
YZ F.S. Rayleigh	$5.00 \times 10^{14} A^2 LK_A$	1.00
YZ Met. Rayleigh	$2.92 \times 10^{14} A^2 LK_A$	1.31

and adjust C for other surface waves to get the same power. K_A is the acoustic wavenumber and all waves are assumed to have the same wavenumber. The 2 factor indicates that the O.F. is a time average. A is the peak amplitude of the surface wave.

C. Overlap Factor Results

Figures 2, 3 and 4 show the overlap factors for the 64° leaky waves, free surface Rayleigh waves, and metallized surface Rayleigh waves, respectively. These curves represent the sum of the electro-optic and elasto-optic contributions. The 64° leaky wave - TM interaction is noteworthy in that its magnitude exceeds all others at the crystal surface. The 41° leaky wave overlap factors are very small and therefore not shown. A thin guide which propagates TM modes and keeps most of the optical energy near the crystal surface should exhibit faster depletion of the incident optical wave with the 64° leaky wave than with any of the other acoustic waves considered.

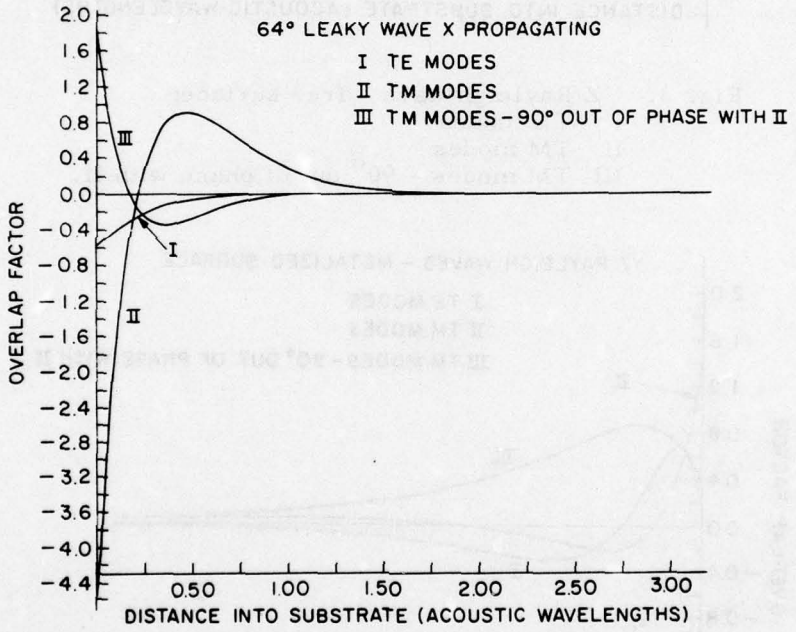


Fig. 2. 64° Leaky wave X propagating:
I TE modes
II TM modes
III TM modes - 90° out of phase with II.

D. Coupling Coefficient Calculations

Using the previous results we investigated coupling coefficient behavior as a function of surface wave type and optical waveguide effective thickness. We assumed a typical graded index profile for the optical waveguide. The profile was approximately

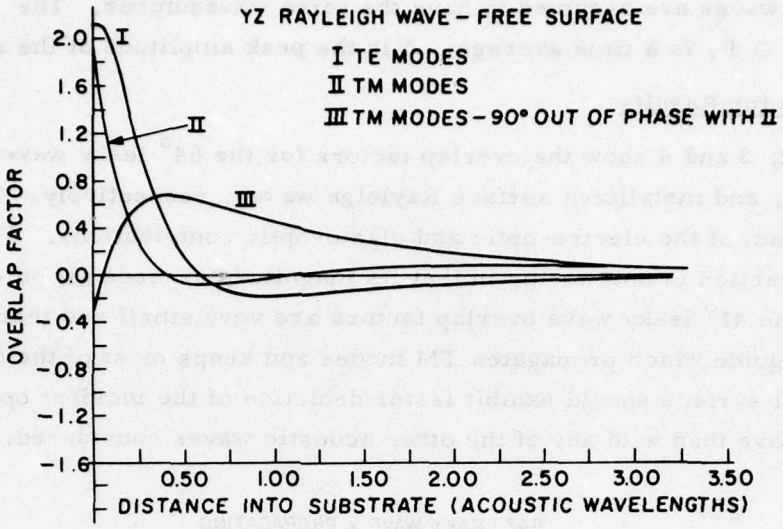


Fig. 3. YZ Rayleigh wave - free surface:
I TE modes
II TM modes
III TM modes - 90° out of phase with II.

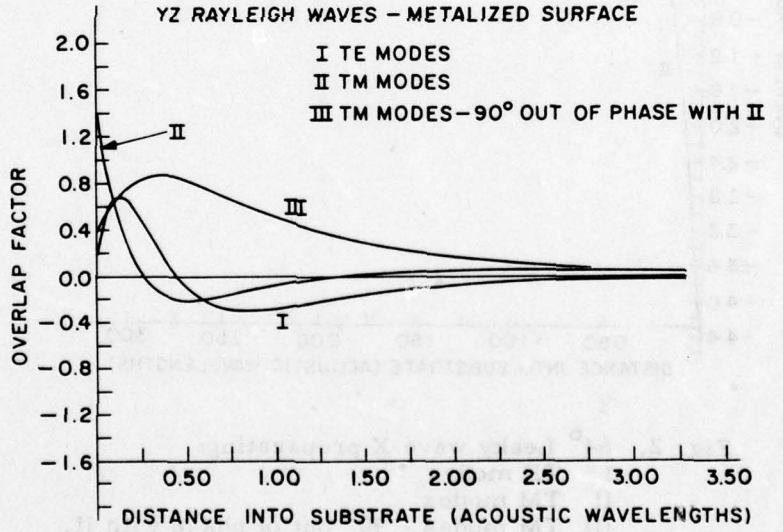


Fig. 4. YZ Rayleigh waves - Metallized surface:
I TE modes
II TM modes
III TM modes - 90° out of phase with II.

complementary error function in shape and therefore similar to profiles reportedly attained by metal indiffusion.[3] Then using a piecewise linear approximation and using Airy function solutions[4] we calculated for TE propagating modes, E(y) and for TM propagating modes, H(y). Figure 5 shows TE 0 and TM 1 waveforms for metallized

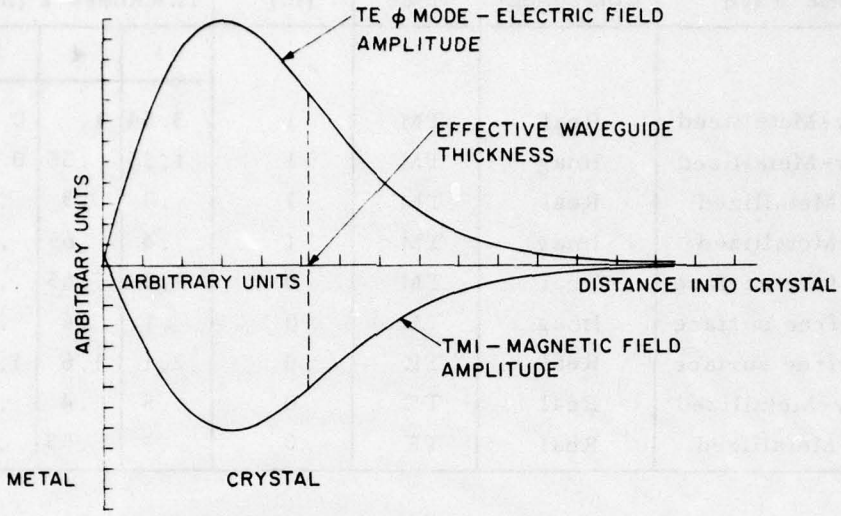


Fig. 5. Electric and magnetic field distributions (not normalized) for typical planar metallized graded index waveguide.

guides. Using these and free surface waveforms and the overlap factor curves we calculate $|\Gamma_{00}|$ *for TE free surface and metallized guides* and TM free surface guides and $|\Gamma_{11}|$ for TM metallized guides. We "compressed" or "stretched" E(y) and H(y) to imitate the effects of different effective waveguide thicknesses and used a numerical integration method for the calculations.

Table I shows that the largest coupling coefficients are obtained with the 64° leaky wave and TM modes and very thin guides, and for the YZ free surface Rayleigh wave and TE modes, also with a thin guide. For thick guides (thicker than 4 microns) we see that Rayleigh-Metallized-TM, Rayleigh-free surface-TM and TE combinations are best. These results are summarized in Table II.

In conclusion, we have found through a theoretical analysis that the 64° rotated Y cut leaky surface wave offers the strongest acousto-optic interaction known to date.

Joint Services Technical Advisory Committee
F44620-74-C-0056

F. Freyre and W-C. Wang

National Science Foundation
ENG76-18901

AD-A050 265 POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 20/3
PROGRESS REPORT NO. 42, 15 SEPTEMBER 1976 - 14 SEPTEMBER 1977, --ETC(U)
NOV 77 A A OLINER F44620-74-C-0056
UNCLASSIFIED POLY-MRI-452.42-77 NL

2 OF 7
AD
A050 265

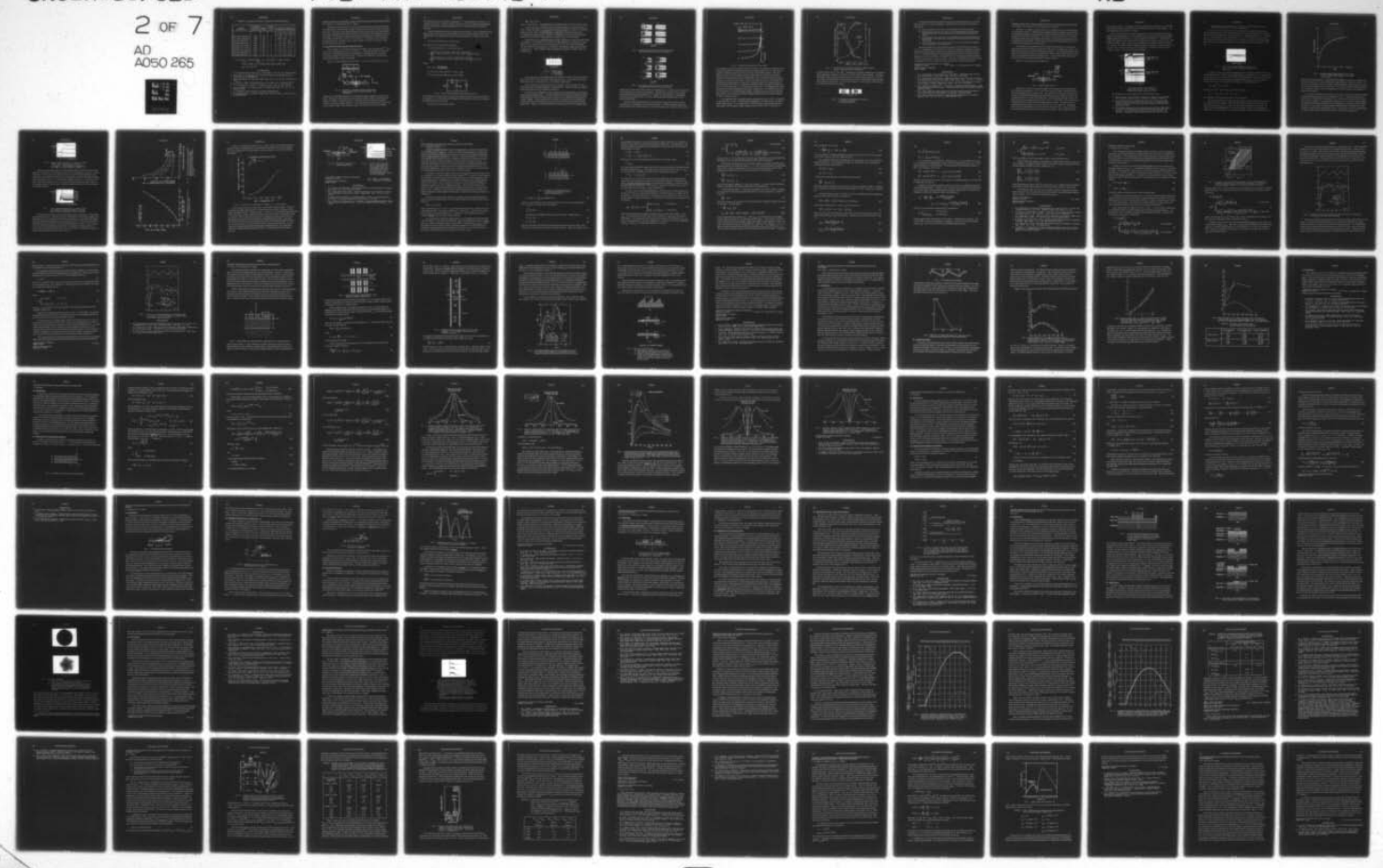


TABLE II. Coupling coefficients vs. effective waveguide thickness.

Surface Acoustic Wave	Coupling Coefficient Component	Optical Mode	Mode Number (m)	Effective Waveguide Thickness T (Microns)*			
				1	4	7	10
64° Leaky-Metallized	Real	TM	1	3.34	1.	0	.4
64° Leaky-Metallized	Imag	TM	1	1.25	.35	0	.2
Rayleigh-Metallized	Real	TM	1	.9	.3	.05	.1
Rayleigh-Metallized	Imag	TM	1	.4	.65	.75	.75
Rayleigh-free surface	Real	TM	0	1.5	.65	.3	.15
Rayleigh-free surface	Imag	TM	0	.1	.4	.5	.6
Rayleigh-free surface	Real	TE	0	2.1	1.6	1.1	.7
64° Leaky-Metallized	Real	TE	0	.5	.4	.3	.2
Rayleigh-Metallized	Real	TE	0	.4	.55	.35	.2

*$T = \lambda_o |\Gamma_{mm}| \; 1.58 \times 10^7 \; \frac{L}{PK_A}$; $K_A = 3.57 \times 10^5 \; m^{-1}$ where P is the power in watts, L the acoustic beam width in meters and $K_A = 3.57 \times 10^5 \; m^{-1}$.

REFERENCES

1. K. Yamanouchi and K. Shibayama, J. Appl. Phys., Vol. 43, No. 3 (March 1972).
2. E.G. Lean, P.F. Heidrich, J.M. White, 1974 Ultrasonics Symposium Proc., IEEE Cat. No. 74 CHO 896-ISU.
3. R.V. Schmidt and I.P. Kaminow, Appl. Phys. Lett., Vol. 25, No. 8 (October 15, 1974).
4. D. Marcuse, IEEE J. Quantum Electron., Vol. QE-9, No. 10 (October 1973).
5. T.C. Lim and G.W. Farnell, J. Acoust. Soc. Am., Vol. 45, No. 4 (1969).
6. J.R. Carruthers, I.P. Kaminow and L.W. Stuly, Appl. Opt., Vol. 13, No. 10 (October 1974).
7. K. Yamanouchi and K. Shibayama, private communication.
8. R.N. Spaight and G.G. Koerber, IEEE Trans. Sonics Ultrason., Vol. SU-18, No. 4 (October 1971).

OPTICAL IMAGING AND MEMORY USING A SILICON DIODE STORAGE CORRELATOR

W-C. Wang, L.S. Rosenheck and K.C. Whang

The optical characteristics (in visible and infrared frequencies) and the memory characteristics of a silicon diode storage correlator have been under study for the past few years. It has been shown experimentally that the storage time is strongly dependent on light intensity and that the storage time in the dark at room temperature is about 40 seconds. We found the storage and writing time constants to be in general agreement with the measured p-n junction characteristics. In addition we have found that the amplitude of the correlation signal, after repetitive writing, increases to a peak and then decreases with time.

A. Description of the Writing and Readout Process

Various types of acoustoelectric storage correlators have been reported.[1-4] The diode array type appears to be the most promising. Although significant inroads have been made in understanding the basic principles of the storage process, due to the structural complexity of the diode mosaic, a detailed interpretation of the observed experimental results is still not complete.

Figure 1 is the experimental configuration. On the $LiNbO_3$ substrate, a pair of

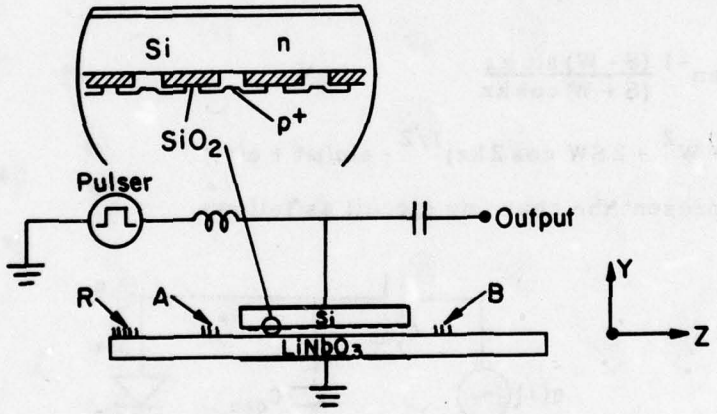


Fig. 1. Schematic of the experimental configuration. Center frequencies of A and B 50 MHz, and R 100 MHz.

50 MHz and one 100 MHz transducers are deposited at their respective locations A, B and R. RCA 4532 vidicon diodes are placed on the $LiNbO_3$ substrate separated by an airgap of ≈ 1000 Å. As depicted in Fig. 1 the diode mosaic is placed such that the diodes are facing the $LiNbO_3$ substrate. The reverse side of the mosaic is coated with a metal and the output of the storage correlator is taken across the Si and $LiNbO_3$.

The writing process is achieved by applying two r.f. pulsed signals of frequency ω at transducers A and B. Through nonlinear interaction of the two oppositely propagating waves, a signal which is time independent but with a spatial variation $\cos(2kz)$ is produced.

We can represent the charge storage process as follows: The two oppositely propagating acoustic waves can be represented as $E_1 = S\sin(\omega t - kz)$ and $E_2 = W\sin(\omega t + kz)$ respectively. The superposition of these two waves generates a stationary wave pattern

$$E_1 + E_2 = E = S\sin(\omega t - kz) + W\sin(\omega t + kz)$$

which can in turn be represented as follows

$$E = (S + W)\cos kz \sin\omega t - (S - W)\sin kz \cos\omega t$$

$$= \left(\sqrt{(S^2 + 2SW + W^2)\cos^2 kz + (S^2 - 2SW + W^2)\sin^2 kz}\right)\sin(\omega t + \phi(z))$$

$$= \left(\sqrt{(S^2 + W^2)(\cos^2 kz + \sin^2 kz) + 2SW(\cos^2 kz - \sin^2 kz)}\right)\sin(\omega t + \phi(z))$$

where

$$\phi(z) = -\tan^{-1}\frac{(S - W)\sin kz}{(S + W)\cos kz}$$

$$E = (S^2 + W^2 + 2SW\cos 2kz)^{1/2} \cdot \sin(\omega t + \phi(z))$$

If we represent the charging circuit as follows

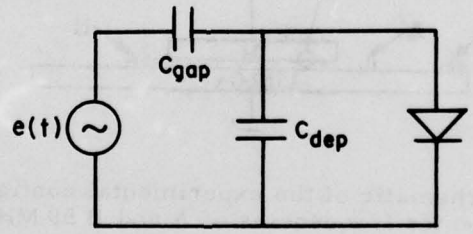


Fig. 2. Circuit representation of the charging process.

we note that the air gap capacitor couples the superposed signal to the semiconductor. The positive r.f. cycles will quickly charge the two capacitors so that one can assume that peak detection is occurring.

We can get a stored charge profile

$$Q(z) = q(z) \cos 2kz$$

with a π/K periodicity. The stored charge will in turn reverse bias the diode system.

The stored charge profile can now be read out by applying a signal $R(t)\cos(2\omega t - 2kz)$. If $R(t) \approx \delta(t)$ then e_{out} is approximately the correlation of S and W. It should be noted, however, that a time scaling factor of one-half is introduced in this type of reading and writing process. This time scaling factor can be removed by using a layered structure BGO-Si-diode-$LiNbO_3$ as pointed out in a previous paper.[5]

Output signals are observed as shown in Fig. 3 where the triangular shaped signal, indicated by I, corresponds to the convolution output between two rectangular writing pulses, both of 2 μsec duration and 20 volts (p-p, across a 50 Ω load) in amplitude. The other pulse of large amplitude corresponds to the applied readout pulse indicated through the air. The correlation output, designated at position II, appears after the readout pulse.

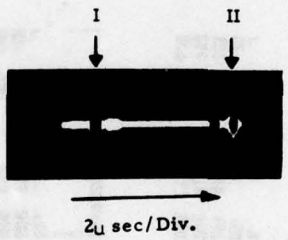


Fig. 3. Signal outputs.
I - Convolution
II - Correlation

If at time t_1 both of the writing pulses are suddenly switched off, but the readout pulse still remains, the correlation output will first rise to a peak and then decrease with time to zero as indicated in Fig. 4, at different light levels.

The storage time in total darkness is ≈ 25 seconds. One can further note that the storage time is strongly dependent on light intensity. Similar results are obtained using an infrared source as indicated in Figure 5.

A reason for the dependence of the storage time on light intensity can be hypothesized in the following manner. The stored charge is discharged through the reversed biased leakage current of the diode array, which in turn is dependent on the minority carrier lifetime. As the light intensity increases, the minority carrier lifetime decreases causing the reversed bias current to increase and in turn the stored charge leaks faster. Figure 6 is the measured diode characteristic which was obtained from

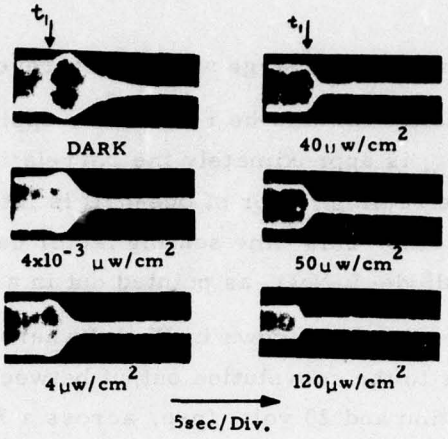


Fig. 4. Correlation output as a function of time with light intensities (visible) as a parameter.

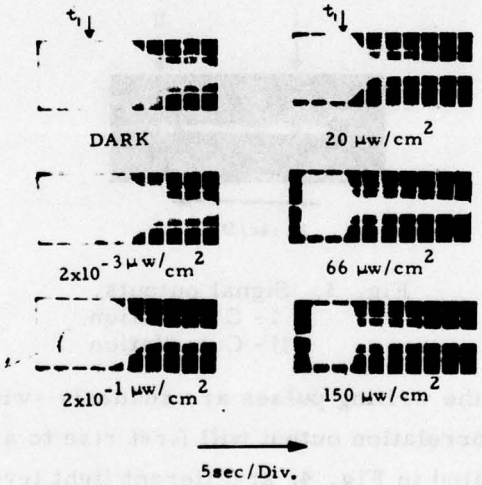


Fig. 5. Correlation output as a function of time with light intensities (infrared) as a parameter.

a diode mosaic with ohmic contacts which are deposited over a surface area of $(2.5\,mm)^2$. A 30 pf capacitance is measured in darkness. The RC time constant is ≈ 20 sec which shows good agreement with storage time. Judging from the sensitivity of the leakage current to very low light levels, it is of no surprise that the storage time varies at very low light levels.

The observed amplitude profile of the correlation output (as a function of time) can perhaps be comprehended in the following manner. Considering the geometry and physical dimensions of the diode arrays, the diode actually only occupies a small portion

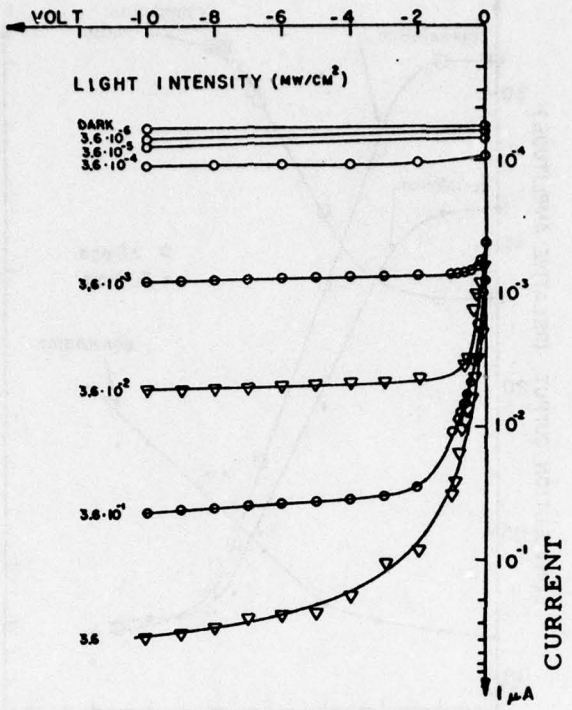


Fig. 6. Reverse characteristics of the diode.

of the surface area ($\approx 1/8$) the rest of the surface area is occupied by an MOS structure. In a previous paper[6] we pointed out that the convolution output in such a structure can actually be predicted by the model of a reversely biased MOS structure. Here the diode serves as a carrier injection and the injected carriers which are the stored charges reverse bias both the diode and the MOS. When the stored charge decreases with time, so does the reverse bias. Since the nonlinear processes of convolution and correlation are quite similar, (except that the stored charge to be correlated with is also decreasing with time), the correlation amplitude profile may exhibit a bump as in the case of the convolution in a reverse biased diode mosaic. The disappearance of the bump when the writing amplitude is decreased, would be due to the simple reason that the reversed bias is too low to cause a bump. A similar reason can be given when the light intensity is increased, the charge storage is decreased and the bump disappears as shown in Figures 4 and 5.

The amplitude of the correlation output as observed in Figs. 3 and 4, is a function of light intensity. In Fig. 7 we plot both the amplitudes of convolution and correlation versus light intensity. It is noted that the correlation output varies linearly through a large range of the variation of light intensity from low levels (4×10^{-6} mw/cm^2) to

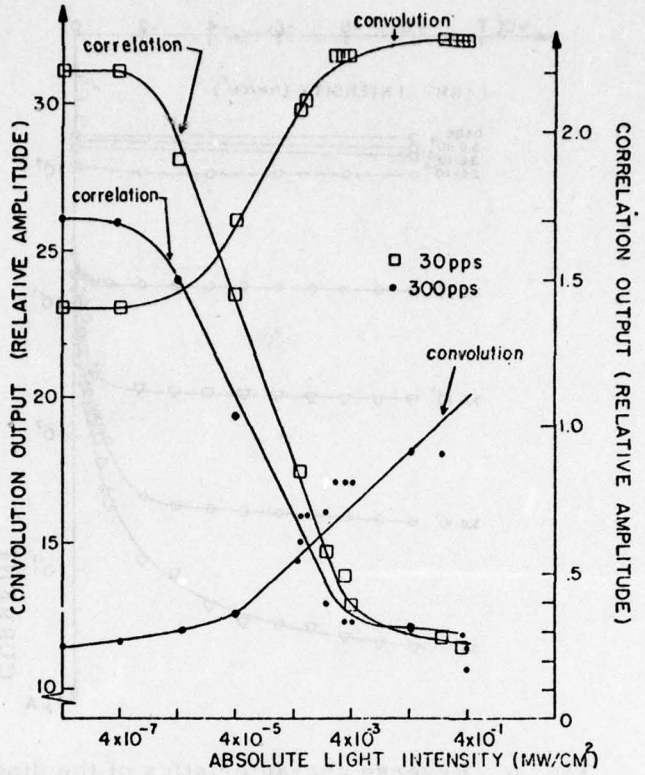


Fig. 7. Correlation and convolution output vs. light intensity with the repetition rate as a parameter.

moderate values (4×10^{-3} mw/cm^2). Although the convolution outputs have a comparable linear range the correlation is sensitive to a lower level of light intensity. The light sensitivity of the storage correlator can be put to use for optical imaging. A narrow slit of light can be projected onto the diode matrix and the correlation output can detect and resolve the slit of light. As can be seen in Fig. 8, a slit of light ≈ 1 mm wide can be resolved by acoustical scanning.



Fig. 8. (a) Acoustic imaging using correlation without slit of light.
(b) With slit of light.

The theoretical model for the physical processes involved with the p-n junction diode correlator is under investigation. This model will be based on the following assumptions

(1) Carrier injection occurs only in the direction of the surface normal.

(2) Diffusion of the injected carriers is only one-dimensional in the direction of the surface wave, x, and confined to the region between two adjacent diodes.

(3) The injection in the y-direction is much faster than the x-directed diffusion-drift process. This allows us to consider the two processes as occurring sequentially.

(4) The recombination time of the minority carriers is related to the leakage current, therefore, the storage time.

We first establish the initial conditions at time $t = 0$ for the device in Figure 1. A $t = 0^+$, a displacement current pulse D of width T is applied to transducers A and B. Using the continuity equation with Poisson's equation the charge distribution is found at $t = T$. The charge distributes through diffusion and drift and recombines through leakage across the p-n junction. The spatial Fourier components of the charge distribution are calculated as a function of time. After this is completed we hope to develop a comprehensive circuit model which will simulate the charge storage process.

Joint Services Technical Advisory Committee
F44620-74-C-0056 W-C. Wang, L.S. Rosenheck and K.C. Whang

National Science Foundation
ENG 76-18901

REFERENCES

1. K.A. Ingebrigtsen, R.A. Cohen and R.W. Mountain, "A Schottky Diode Acoustic Memory Correlator," Appl. Phys. Lett., Vol. 26, p. 596 (1974).

2. K.A. Ingelbrigtsen and E. Stern, "Coherent Integration and Correlation in a Modified Acoustoelectric Memory Correlator," Appl. Phys. Lett., Vol. 27, p. 170 (1975).

3. C. Maerfeld, Ph. Defranould and P. Tournois, "Acoustic Storage and Processing Device Using p-n Diodes," Appl. Phys. Lett., Vol. 27, p. 577 (1975).

4. W-C. Wang, K.C. Whang and L.S. Rosenheck, "Optical Imaging and Memory Using a Silicon Diode Storage Correlator," Ultrasonics Symposium Proceedings, Annapolis, Maryland (1976).

5. W-C. Wang, "SAW Correlator Using Layered Structure $Bi_{12}GeO_{20}$-Si-$LiNbO_3$," Ultrasonics Symposium Proceedings, Los Angeles, California (1975).

6. W-C. Wang, "Physical Processes of SAW Schottky-Diode Memory Correlator," Electron. Lett., Vol. 12, No. 4, p. 105 (February 1976).

CHARGE STORAGE IN A DIODE MOSAIC INDUCED BY SURFACE ACOUSTIC WAVES

W-C. Wang, L.S. Rosenheck, K.C. Whang and H. Schachter

Research on storage correlators has been an ongoing project by various groups for the past few years.[1-4] Although these investigators have proposed various models to represent the charge storage process, to date a completely cohesive theory for the storage mechanism has not emerged. Reasons for this are twofold: first, different writing processes have been employed and second, the geometry of the diodes and the storage devices are different. These devices range from Schottky diodes to p-n junction diodes with neighboring MOS capacitors.

To facilitate the understanding of the charge storage process we conducted an experiment to monitor the total charge in the system due to a single sonic pulse, multiple sonic pulses, and contra-directional sonic pulses. From the experimental results we also learned that a transverse acoustoelectric effect which accompanies the surface wave will contribute during the charging process. The acoustoelectric effect plays a role in forward and reverse biasing the p-n junction diodes.

Figure 1 is the experimental configuration.

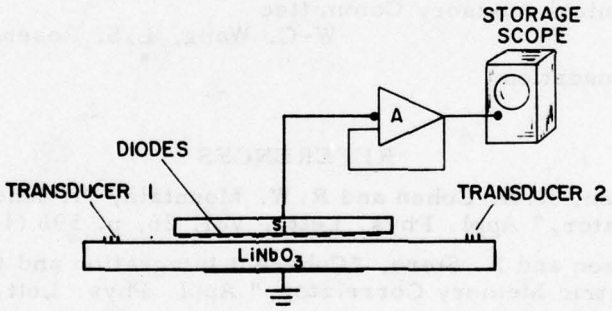


Fig. 1. Experimental set-up.

We detect our slowly time varying output by first entering a unity gain op amp with a high input impedance, $>10^{12}\Omega$, and a slew rate of $70^{V}/\mu sec.$ and the op amp output is then monitored by a storage scope. The op amp acts to make the output of the storage correlator load independent and the storage scope is necessary to monitor the charge build-up due to just a single SAW pulse. On the $LiNbO_3$ substrate a pair of 50 MHz transducers are deposited at locations 1 and 2. An RCA 4532 vidicon diode mosaic is placed above the $LiNbO_3$ substrate supported by an airgap. The diode mosaic is manufactured by first using an n type Si wafer of 50 Ω-cm resistivity. A thin SiO_2 layer with $\approx 5\,\mu m$ diameter holes is first formed on the Si surface. Boron impurities are then diffused into the hole site to form the p-n junction. The periodicity of the

diode array is ≈ 12 μm. The diodes are then overlayed with a p^+-Si pad. An antireflection coating (SiO) of several Å thick is also deposited at the back of the Si wafer. The total thickness of the Si wafer is ≈ 12 μm.

We are interested in the charge storage which is due to a single acoustic pulse applied to just one transducer. Figures 2(a) and (b) show typical results of the output signals generated by successive r.f. pulses applied to transducer 1. The difference between the figures is that in Fig. 2(a) the pulse width is 13 μsec. whereas in Fig. 2(b) the pulse width is 6 μsec. In each of the oscillograms we labelled the first 5 traces 1-5. Trace 1 was obtained when we applied a pulse to transducer 1 with the system uncharged. Each of the successive pulses is applied at .2 second intervals and each respective output is indicated by traces 2 through 5.

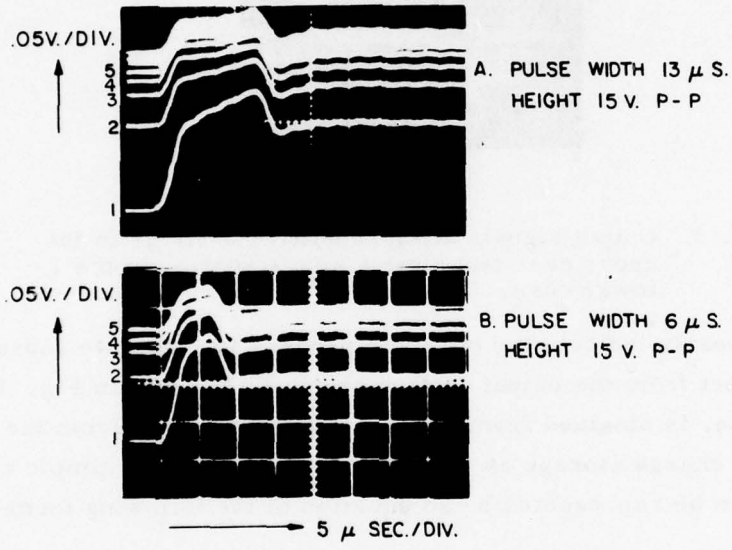


Fig. 2. Output signal generated by repetitive single pulse writing. The time interval between successive pulses is 0.2 sec.

An individual trace can be understood in the following manner:

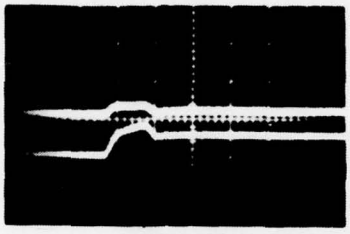
(1) When the pulse is first applied to transducer 1 but prior to the propagation of the pulse under the semi-conductor no voltage is observed.

(2) As the pulse begins to propagate under the semi-conductor the output voltage increases rapidly until the acoustic pulse is completely under the semi-conductor at which point the output voltage increases but at a much slower rate.

(3) As the pulse emerges from under the semi-conductor the output voltage decreases until the pulse has completely exited from under the semi-conductor. The output voltage is then just the charge storage. We see that as we apply successive pulses the next trace starts at about

the previous charge storage level, however, the increment increase in output voltage is less on each successive pulse.

In Fig. 2(a) it can be seen that after the fifth pulse was applied the successive pulses contributed little to the charge storage level. The observed variation in the output voltage is then totally due to the transverse d.c. acoustoelectric effect. To indicate the d.c. acoustoelectric effect more clearly, in Fig. 3 we compare the output signal due to just a single writing and that which occurs in the same system after many repetitive writings. For the latter case the d.c. acoustoelectric effect is symmetric and flat.



OUTPUT SIGNAL

Fig. 3. Output signals after repetitive writings in the upper case and after a single writing in the lower case.

In order to clearly indicate the charging process vs. time we substracted the d.c. acoustoelectric effect from the output voltage profile. The plot in Fig. 4, charge storage as a function of time, is obtained from trace 1 of Figure 2(a). From the curve we know that the build-up of charge storage as a function of time is not a simple exponential. In fact the curve can be represented by an equation of the following form:

$$K_3 - (K_2 e^{-t/\tau_2} - K_1 e^{-t/\tau_1})$$

where $\tau_2 = 4.2 \times 10^{-6}$ secs. and $\tau_1 = 3.5 \times 10^{-6}$ seconds

$K_3 = 2$ $K_2 = 5.8$ and $K_1 = 3.9$.

An explanation of the charging process is now presented.

Because the diodes are isolated from each other only those diodes directly above the acoustic wave are charged. The respective diode is charged to or near its full charge value within a few r.f. cycles, but, the remaining diodes act as a load. As more diodes become charged they can be considered as voltage sources loaded by a capacitive and resistive load. For example, consider the case of 1,000 diodes in the

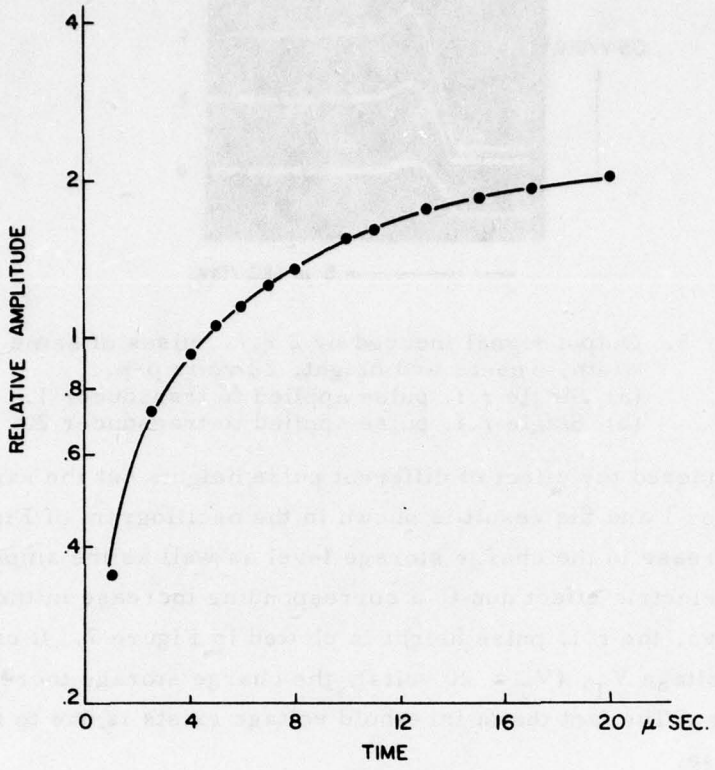


Fig. 4. Charge storage (single writing) as a function of time. Pulse width 13 μsec., pulse height 15 volts p-p.

direction of SAW propagation. If initially the first diode is charged to a voltage V and the remaining diodes are uncharged then the observed output voltage is V/1000. If n diodes are charged the output is nV/1000. Thus, the output voltage represents a measure of the total charged stored in the system. If one considers the storage mechanism for an individual diode, it can be represented by a simple voltage source applied to a diode and a capacitor.

To further illustrate the accumulation of charge in the system consider the effect of applying pulses simultaneously to transducers 1 and 2. Using the same structure as in Fig. 2 we obtained the oscillogram of Fig. 5 where each trace starts at 0 volts. In the case of a single r.f. pulse applied to transducer 1 or a single r.f. pulse applied to transducer 2 the output voltages from the respective traces ((a) and (b)) are nearly the same. For the case of the pulses applied simultaneously to transducers 1 and 2 we have a result which is indicated by trace (c). Notice that trace (c) is approximately the additive sum of trace (a) and trace (b) which confirms our explanation.

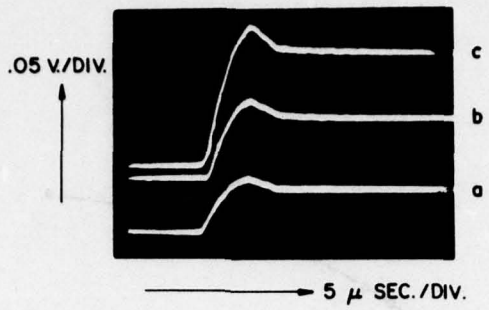


Fig. 5. Output signal induced by 2 r.f. pulses of same width, 6 μsec. and height, 22 volts p-p.
(a) Single r.f. pulse applied to transducer 1.
(b) Single r.f. pulse applied to transducer 2.

We also considered the effect of different pulse heights but the same pulse width applied to transducer 1 and the result is shown in the oscillogram of Figure 6. We note that there is an increase in the charge storage level as well as the amplitude of the transverse acoustoelectric effect due to a corresponding increase in the applied voltage. The stored charge vs. the r.f. pulse height is plotted in Figure 7. It can be seen that after a threshold voltage V_T, ($V_T \approx 20$ volts), the charge storage increase linearly with applied voltage. The fact that a threshold voltage exists is due to the nonlinearity of the diode response.

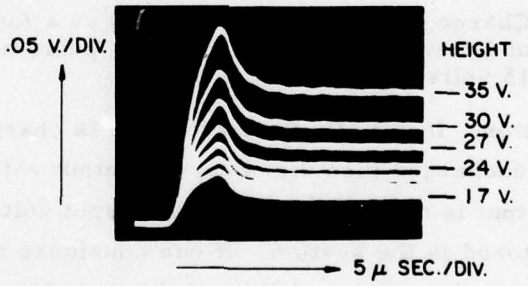


Fig. 6. Output signals generated by r.f. input of 6 μsec. pulse width but of different heights. The time interval between the pulses is over 30 seconds.

Charge storage will also increase with respect to an increasing r.f. pulse width until saturation sets in. Saturation occurs in time because as the airgap, depletion and MOS capacitors are charged toward a particular voltage the current due to the applied pulse decreases until eventually the current is equal to zero. At this point there is no additional charge accumulation as is illustrated effectively in Figure 8. The variation of incremental charge deposition for the case of 5 repetitive writings with the r.f. pulse width as a parameter is indicated. We can see that the incremental charge storage decreases dramatically after the first write in.

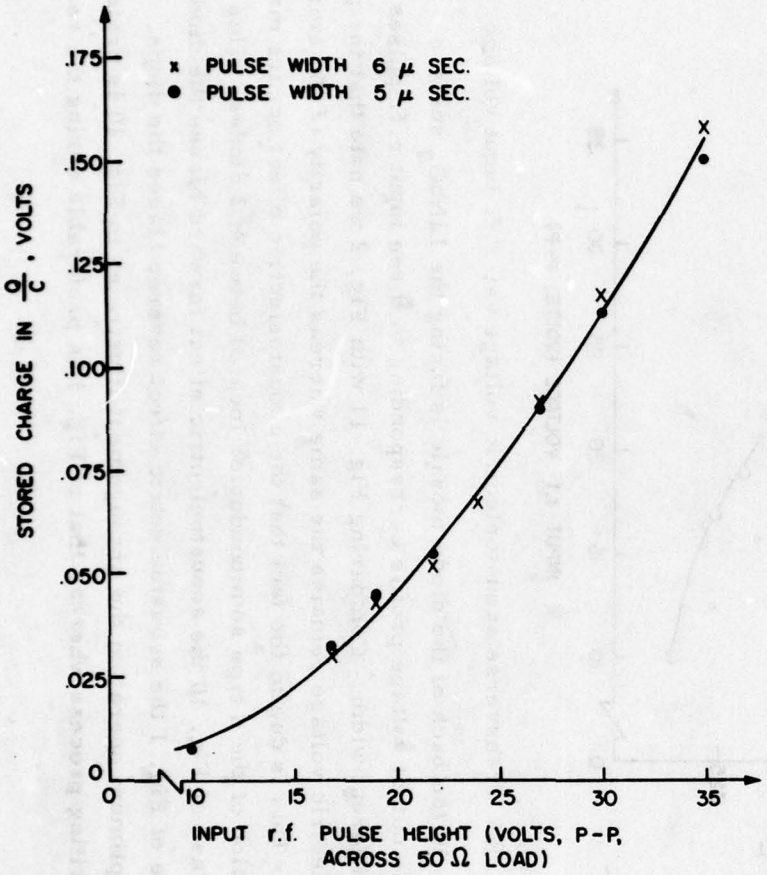


Fig. 7. Charge storage due to single writing vs. r.f. input pulse height.

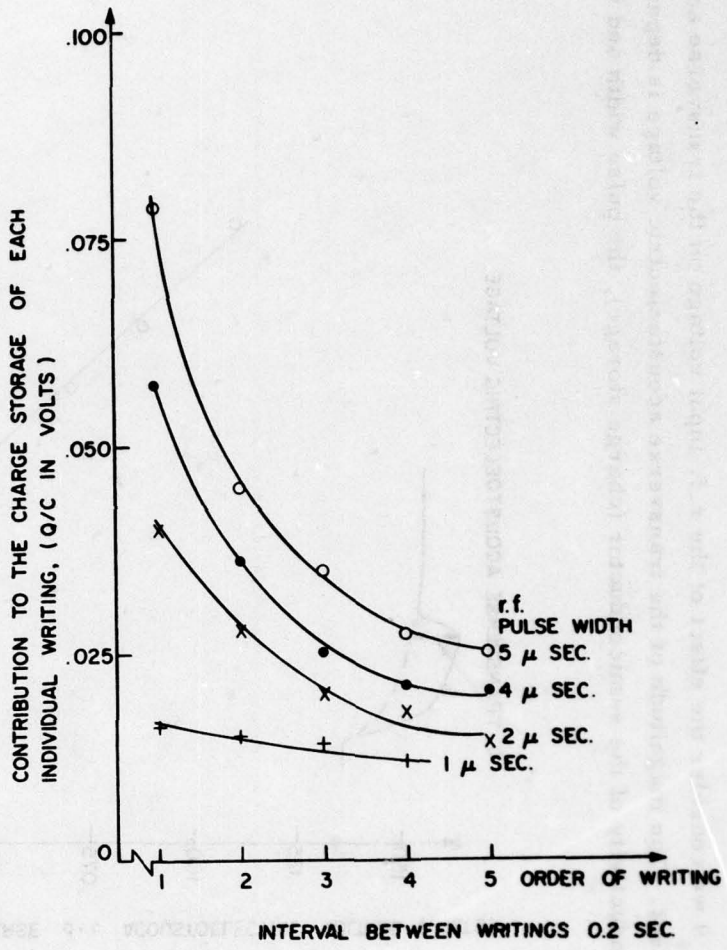


Fig. 8. Variation of incremental charge deposition of each individual writing with pulse width as a parameter. The time interval between writings is 0.2 sec.

In Fig. 9 we consider the effect of the r.f. input voltage on the transverse acoustoelectric effect. The magnitude of the transverse acoustoelectric voltage is dependent on the conductivity of the semiconductor (charge storage), the pulse width and the pulse height.

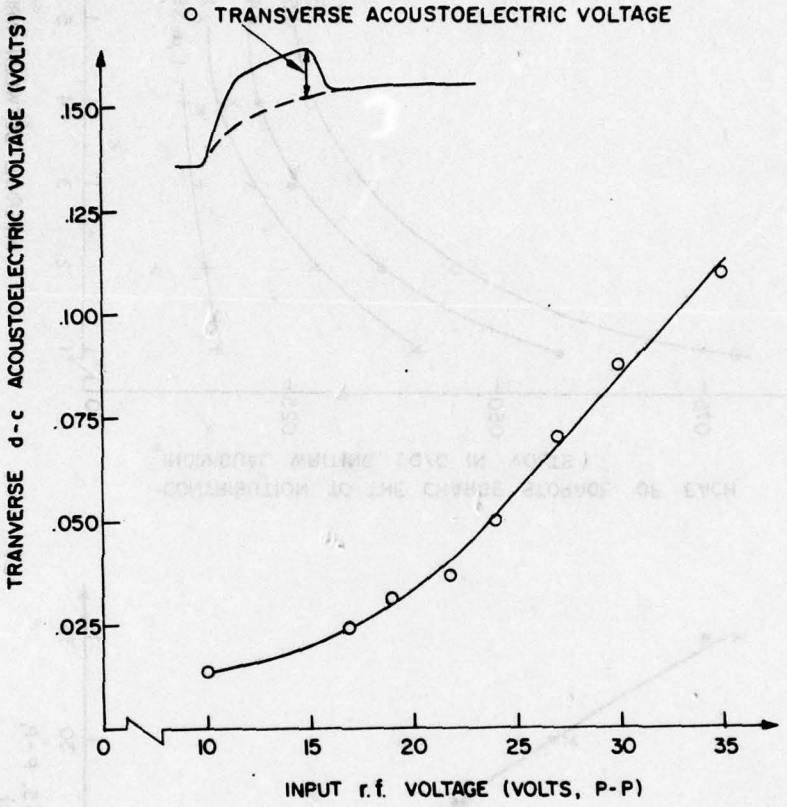


Fig. 9. Transverse acoustoelectric voltage vs. r.f. input voltage.

In Fig. 10 the back of the diode mosaic is facing the $LiNbO_3$ surface. Figure 11 illustrates the output voltage profile corresponding to three input r.f. pulses of equal height, but different width. Comparing Fig. 11 with Fig. 2 we note that the polarity of the the acoustoelectric voltage remains the same whereas the polarity of the stored charge is reversed. This is due to the fact that the acoustoelectric effect occurs mainly inside that portion of the n type semiconductor located between 2 diodes. This shows that for the case of Fig. 10 the acoustoelectric effect forward biases the diodes whereas in the case of Fig. 1 the acoustoelectric effect reverse biases the diode. In conclusion for optimum operation the arrangement illustrated in Fig. 10 is preferable during the writing process whereas that of Fig. 1 is preferable during the readout process.

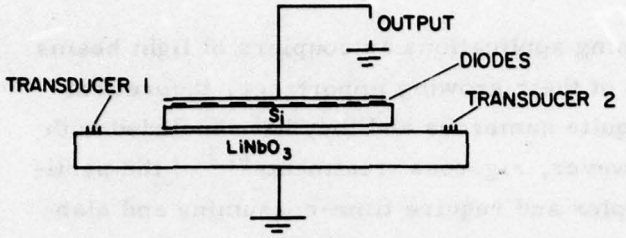


Fig. 10. Modified configuration of original experiment.

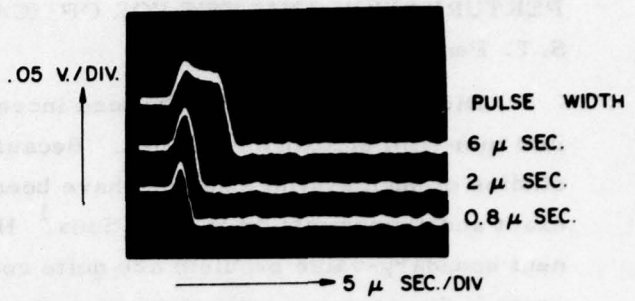


Fig. 11. Output signals generated by r.f. pulses with different pulse widths and a pulse height of 30 volts p-p (back of diode array facing $LiNbO_3$). The time interval between pulses is over 30 seconds.

Joint Services Technical Advisory Committee
F44620-74-C-0056

National Science Foundation
ENG76-18901

W-C. Wang, L.S. Rosenheck
K.C. Whang and H. Schachter

REFERENCES

1. A. Bers and J.H. Cafarella, "Surface State Memory in Acoustoelectric Correlator," Appl. Phys. Lett., pp. 25 and 133 (1974).

2. K.A. Ingebrigtsen and E. Stern, "Coherent Integration and Correlation in a Modified Acoustoelectric Memory Correlator," Appl. Phys. Lett., Vol. 27, p. 170 (1975).

3. Ph. Defranould, H. Gautier, C. Maerfeld and P. Tournois, "P-N Diode Memory Correlator," Ultrasonics Symposium Proceedings, p. 336 (1976).

4. W-C. Wang, K.C. Whang and L.S. Rosenheck, "Optical Imaging and Memory Using a Silicon Diode Storage Correlator," Ultrasonics Symposium Proceedings, p. 352 (1976).

PERTURBATION ANALYSIS FOR OPTICAL GRATING COUPLERS

S. T. Peng and T. Tamir

Dielectric gratings have found increasing applications as couplers of light beams into thin-film optical waveguides. Because of their growing importance, theoretical studies of such grating couplers have been quite numerous and they have included both exact and approximation formulations.[1] However, rigorous treatments[2, 3] of the pertinent boundary-value problem are quite complex and require time-consuming and elaborate high-precision computer programs to yield accurate quantitative results. To avoid this, perturbation procedures have been proposed[1] for obtaining approximate results, but most of these are either restricted to grating having shallow grooves or to specialized applications. The results so far available have therefore not been adequate to develop systematic criteria for the design of grating couplers.

In the past, we have developed an improved perturbation procedure that was shown to be suitable for an accurate analysis of beam couplers.[4, 5] In contrast to other methods, this approach is not restricted to shallow grating grooves and it can be applied to arbitrary grating profiles. Furthermore, the analytic formulation of the electromagnetic fields can be viewed in terms of equivalent transmission-line networks that lend considerable insight into the wave-coupling mechanism. These networks enable one to assess the role of each grating parameter, as discussed separately in this report (see Reference 6). In addition, we have observed that the previous analytic results can be further simplified to obtain explicit formulas that are particularly useful for design considerations.

To carry out a perturbation analysis of a dielectric grating waveguide having an arbitrary profile as shown in Fig. 1(a), the relative dielectric constant is written everywhere as

$$\varepsilon(x, z) = \varepsilon_u + \varepsilon_p(x, z) \tag{1}$$

where the subscript $u = s, f, r$ and a in ε_u denotes the substrate, film, residual layer and superstrate (usually, air) regions, respectively. The term $\varepsilon_p = \varepsilon_p(x, z)$ is zero everywhere except in the grating region ($0 < z < t_g$) where $u = g$ and ε_g then refers to the (volume) <u>average</u> value of $\varepsilon(x, z)$ inside that region.

If ε_p vanishes also inside the grating region, we are dealing with a multi-layered uniform medium. This multi-layered medium, which is regarded henceforth as a <u>basic</u> structure, supports well-known surface-wave modes. The quantity $\varepsilon_p = \varepsilon_p(x, z)$, which describes the actual periodic variation inside the grating layer, is then viewed as a perturbation imposed on that basic (unperturbed) structure. Due to the periodicity of the grating, this perturbation can be written as a Fourier series

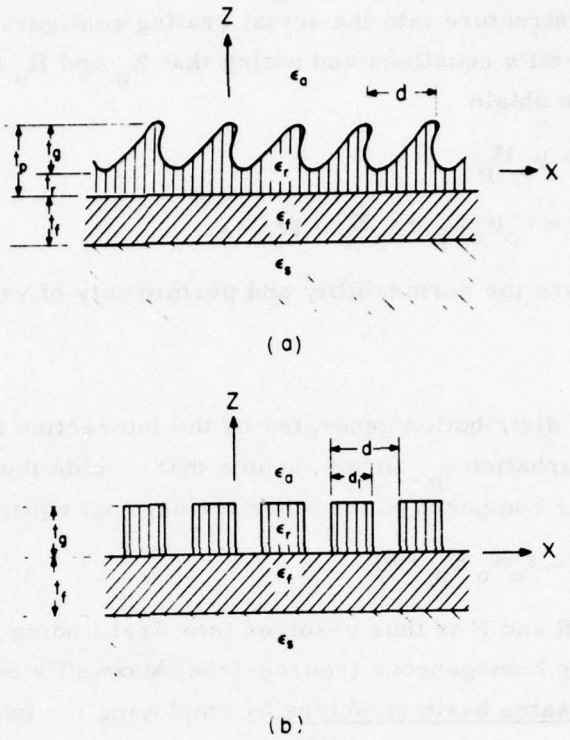


Fig. 1. Geometry of the grating structure:
(a) General asymmetric profile;
(b) Rectangular profile.

$$\epsilon_p = \epsilon_p(x,z) = \sum_{n=-\infty}^{\infty} \epsilon_n(z) \exp(i2n\pi x/d) \quad . \tag{2}$$

Because ϵ_g was defined as the average permittivity inside the entire grating layer $0 < z < t_g$, the n = 0 term in Eq. (2) satisfies

$$\int_0^{t_g} \epsilon_0(z)\, dz = 0 \quad . \tag{3}$$

The electric and magnetic fields may also be written, respectively, as

$$E = E_u + E_p \quad , \tag{4}$$

$$H = H_u + H_p \quad . \tag{5}$$

Here E and H refer to the fields in the actual grating structure shown in Fig. 1(a), whereas E_u and H_u denote the fields in the basic structure. Hence E_p and H_p refer to

perturbations superposed on E_u and H_u when the periodic variation ϵ_p is introduced to modify the basic structure into the actual grating configuration. Introducing Eqs. (4) and (5) into Maxwell's equations and noting that E_u and H_u also satisfy these equations independently, we obtain

$$\nabla \times E_p = i\omega \mu_o H_p \quad , \tag{6}$$

$$\nabla \times H_p = -i\omega \epsilon_o(\epsilon_u E_p + \epsilon_p E_p + p) \quad , \tag{7}$$

where μ_o and ϵ_o are the permeability and permittivity of vacuum, while

$$p = \epsilon_p E_u \tag{8}$$

is a known dipole distribution generated by the interaction between the incident field and the periodic perturbation ϵ_p. If we assume that, inside the brackets of Eq. (7), $\epsilon_p E_p$ is of second order compared to the other terms, that equation reduces to

$$\nabla \times H_p = -i\omega \epsilon_o(\epsilon_u E_p + p) \quad . \tag{9}$$

The solution for E and H is thus resolved into first finding E_u and H_u along the basic structure by using homogeneous (source-free) Maxwell's equations. We then find E_p and H_p along the same basic structure by employing the inhomogeneous (dipole source) Maxwell's Equations (6) and (9). This leads to a result for E and H whose accuracy is affected mainly by the omission of $\epsilon_p E_p$ in Equation (7).

Because of the two-dimensional geometry, the surface-wave modes supported by the basic structure are invariant with respect to y and their fields are of either the TE or TM type. The field solutions of a surface-wave mode in the basic structure can be easily obtained by standard techniques. In particular, the total surface-wave power can be given by:

$$P_{SW} = \int_{-\infty}^{\infty} E_u \times H_u^* \cdot x_o \, dz = \begin{cases} \frac{1}{2}(\beta_0/\omega \mu_o) t_{eff}, & \text{for TE modes} \\ \frac{1}{2}(\beta_0 \omega \epsilon_o \epsilon_f / k_{zf}^2) t_{eff}, & \text{for TM modes} \end{cases} \quad , \tag{10}$$

where the asterisk denotes complex conjugate and x_o is a unit vector along x. Here $t_{eff} > t_f$ is an effective thickness which can be assumed to contain all of the surface-wave power. For the general case of multi-layer structure, the formulae for t_{eff} are rather complicated and are therefore omitted here. However, for the two simple limit cases $t_g = 0$ and $t_g \to \infty$ already discussed above, we have

$$t_{eff} = \begin{cases} t_f + \dfrac{1}{a_s} + \dfrac{1}{a_m} & \text{for TE modes} \\ t_f + \dfrac{1}{a_s} \cdot \dfrac{\varepsilon_f \varepsilon_s (b_f^2 + a_s^2)}{(\varepsilon_f a_s)^2 + (\varepsilon_s b_f)^2} + \dfrac{1}{a_m} \cdot \dfrac{\varepsilon_f \varepsilon_m (b_f^2 + a_m^2)}{(\varepsilon_f a_m)^2 + (\varepsilon_m b_f)^2}, & \text{for TM modes} \end{cases} \tag{11}$$

where m = a or g. It is interesting to note that t_{eff} is identical to the effective thickness that takes into account the Goos-Hanchen shift in thin-film waveguides.[1]

The presence of the grating can modify the surface-wave fields so that they leak energy into the air and substrate regions. For a first-order perturbation, the perturbed (leaky wave) fields are represented in terms of the space-harmonics, each of which independently satisfies the source-excited transmission-line relations:

$$\frac{dV_n}{dz} = ik_{zn} Z_n I_n - v_n \quad , \tag{12}$$

$$\frac{dI_n}{dz} = ik_{zn} Y_n V_n - j_n \quad , \tag{13}$$

where the distributed voltages $v_n = v_n(z)$ and currents $j_n = j_n(z)$ are related to the basic surface-wave fields in the grating region and are therefore known.

The leakage parameter α of the perturbed (leaky) wave can be found by recalling that the power $P = P(x)$ along the perturbed structure varies as $\exp(-2\alpha x)$. We therefore have

$$\frac{dP}{dx} = -2\alpha P \quad , \tag{14}$$

where the change dP in the power is due to those harmonic fields that radiate away from the structure. We then obtain

$$-\frac{dP}{dx} = P_{rad} = \sum_n P_n \quad , \tag{15a}$$

$$P_n = P_n^{(s)} + P_n^{(a)} = |V_n^{(s)}|^2 \, \mathrm{Re}[Y_n^{(s)}] + |V_n^{(a)}|^2 \, \mathrm{Re}[Y_n^{(a)}] \quad , \tag{15b}$$

where Re(q) denotes the real part of q. Here P_{rad} is the total power per unit length (along x) that has leaked into the substrate and air regions. As this power is given by those scattered waves that propagate (rather than decay) in those regions, n in Eq. (15) runs over only those values for which $Y_n^{(a)}$ and/or $Y_n^{(s)}$ are real. Because $\varepsilon_s > \varepsilon_a$, it is easy to verify that sometimes only $p_n^{(s)}$ may contribute to p_n, while $p_n^{(a)}$ is zero, i.e., that n-th diffracted order radiates in the substrate but not in the air region. Combining

Eq. (14) with Eq. (15), we get

$$\alpha = \frac{P_{rad}}{2P} = \sum_n \alpha_n \quad , \quad \text{with} \quad \alpha_n = \frac{P_n}{2P} \quad , \tag{16}$$

so that α_n denotes the (partial) leakage accounted for by the n-th line in which propagation occurs in the air and/or the substrate region.

The coupled transmission line equations, Eq. (12) and Eq. (13), may be solved in many different ways. For the present problem, we find that it is more convenient to introduce the linear transformation:

$$U_n^{\pm} = \frac{1}{2}(V_n \pm Z_n I_n) \quad , \tag{17}$$

$$u_n^{\pm} = \frac{1}{2}(v_n \pm Z_n j_n) \quad , \tag{18}$$

to obtain the two decoupled first order differential equations:

$$\frac{dU_n^{\pm}}{dz} = \pm ik_{zn} U_n^{\pm} - u_n^{\pm} \quad . \tag{19}$$

It may be readily verified from Eq. (19) that, for the time dependence $\exp(-i\omega t)$ adopted here, U_n^+ and U_n^- are voltage waves that travel along the +z and - z directions, respectively.

Inside the grating region, the solution of the two decoupled first-order differential equations in Eq. (19) can readily be written as

$$U_n^{\pm}(z) = [U_n^{\pm}(0) - \int_0^z u_n^{\pm}(z') \exp(\pm ik_{zn} z')] \exp(\pm ik_{zn} z) \quad . \tag{20}$$

The boundary conditions at $z = 0$ and $z = t_g$ can be written as

$$U_n^+(0) = r_{gn}^- U_n^-(0) \quad \text{and} \quad U_n^-(t_g) = r_{gn}^+ U_n^+(t_g) \quad , \tag{21}$$

where r_{gn}^+ and r_{gn}^- are the reflection coefficients of the n-th harmonic at the upper and lower surfaces of the grating, respectively. Introducing Eq. (20) into Eq. (21), we obtain

$$U_n^-(0) = \frac{G_n^+ r_{gn}^+ \exp(i2k_{zn} t_g) - G_n^-}{R_n} \quad , \tag{22}$$

$$U_n^+(t_g) = \frac{(G_n^+ - r_{gn}^- G_n^-) \exp(ik_{zn} t_g)}{R_n} \quad , \tag{23}$$

with

$$G_n^{\pm} = \int_0^{t_g} u_n^{\pm}(z) \exp(\pm ik_{zn} z)dz \quad , \tag{24}$$

$$R_n = 1 - r_{gn}^{+} r_{gn}^{-} \exp(i2k_{zn} t_g) \quad . \tag{25}$$

It is noted that R_n is generally non-vanishing for $n \neq 0$. The condition $R_n \neq 0$ for $n \neq 0$ is equivalent to excluding the resonant case of Bragg reflection $\beta_0 d = m\pi$. Hence $U_n^{-}(0)$ and $U_n^{+}(t_g)$ are well determined. We may therefore find the powers needed in Eq. (15) by using Eq. (21) to yield

$$p_n^{(a)} = \mathrm{Re}[Y_n^{(a)} | V_n^{(a)} |^2] = (1 - | r_{gn}^{+} |^2) | U_n^{+}(t_g) |^2 \mathrm{Re}[Y_n^{(g)}] \quad , \tag{26}$$

$$p_n^{(s)} = \mathrm{Re}[Y_n^{(s)} | V_n^{(s)} |^2] = (1 - | r_{gn}^{-} |^2) | U_n^{-}(0) |^2 \mathrm{Re}[Y_n^{(g)}] \quad . \tag{27}$$

With the values of p_n given by Eqs. (26) and (27), the leakage parameter α can be found via Equations (15) and (16).

In general, the foregoing relations yield α in the form of an elaborate function of all the grating parameters. However, for a grating having a rectangular profile shown in Fig. 1(b), most of the preceding relations simplify considerably. In this case, the integrals for $G_n^{\pm}$ in Eq. (24) can be explicitly evaluated as:

$$G_n^{\pm} = -\frac{\omega \varepsilon_o \varepsilon_n}{2Y_n} C \left\{ (K_n \quad 1) \frac{1 - \exp[i(k_z \pm k_{zn})t_g]}{k_z \pm k_{zn}} \right.$$

$$\left. + (K_n \pm 1) r^{+} \exp(i2k_z t_g) \frac{1 - \exp[-i(k_z \pm k_{zn})t_g]}{k_z \pm k_{zn}} \right\} \quad , \tag{28}$$

where C is a normalization constant and K_n is defined by:

$$K_n = \begin{cases} 0 & \text{for TE modes,} \quad , \\ \beta_0 \beta_n / k_z k_{zn} \quad , & \text{for TM modes} \quad . \end{cases} \tag{29}$$

Inserting $G_n^{\pm}$ into Eqs. (22) and (23), we can find α via Equations (15) and (16). This can be easily carried out by numerical computation. However, for small and large values of t_g, we can derive the following simplified approximations:

$$\alpha_n = \begin{cases} \dfrac{A_n}{4Y_n P}\left[\left(1 + K_n \dfrac{Y_a}{Y_g}\right)\omega\varepsilon_o\varepsilon_n t_g \cos\Psi\right]^2 , & \text{for } t_g \text{ small} \\ \dfrac{B_n}{4Y_n P}\left|\dfrac{K_n + 1}{k_z + k_{zn}}\,\omega\varepsilon_o\varepsilon_n \cos\Phi\right|^2 . & \text{for } t_g \text{ large} \end{cases} \tag{30}$$

where A_n and B_n are both nearly equal to unity for most practical cases.

Another interesting quantity is the partition of energy leaked into the air and substrate regions. The upper and lower bounds for the ratio of power radiated into the air and the substrate regions are given, respectively, by

$$\left.\frac{p_n^{(a)}}{p_n^{(s)}}\right|_{ub} = \frac{1 + |r_n^+|}{1 - |r_n^+|}\,\frac{1 + |r_n^-|}{1 - |r_n^-|} , \tag{31a}$$

$$\left.\frac{p_n^{(a)}}{p_n^{(s)}}\right|_{\ell b} = \frac{1 - |r_n^+|}{1 + |r_n^+|}\,\frac{1 - |r_n^-|}{1 + |r_n^-|} . \tag{31b}$$

Since the geometrical mean of Eq. (31) is unity and $|r_n|$ is usually small, the leakage power tends to divide equally between the substrate and air regions in most cases.

The formulas in Eqs. (30) and (31) are simple enough for computations using a hand-calculator; they are therefore particularly useful for practical design of periodic optical couplers, as we shall show in the following report.

Office of Naval Research
N00014-75-C-0421

S. T. Peng

National Science Foundation
ENG-74-23908

REFERENCES

1. For background and an extensive bibliography up to early 1975, see T. Tamir, Ed., "Integrated Optics," (New York: Springer-Verlag, 1975) Chap. 3, p. 84.
2. N. Neviere, R. Petit and M. Cadilhac, "About the Theory of Optical Grating Coupler-Waveguide Systems," Opt. Commun., Vol. 8, pp. 113-117 (June 1973).
3. S.T. Peng and T. Tamir, "Theory of Periodic Dielectric Waveguides," IEEE Trans. Microwave Theory Tech., Vol. MTT-23, pp. 123-133 (January 1975).
4. K. Handa, S.T. Peng and T. Tamir, "Improved Perturbation Analysis of Dielectric Gratings," Appl. Phys., Vol. 5, pp. 325-328 (January 1975).
5. S.T. Peng and T. Tamir, "TM-Mode Perturbation Analysis of Dielectric Gratings," Appl. Phys., Vol. 7, pp. 35-38 (May 1975).
6. T. Tamir, S.T. Peng and K.C. Chang, "Network Methods for Dielectric-Grating Applications," Progress Report No. 42 to JSTAC, Polytech. Inst. of New York, Report No. R-452.42-77 (1977).

DESIGN OF GRATING COUPLERS

T. Tamir and S. T. Peng

The coupling performance of a grating coupler is known[1] to be most strongly influenced by the leakage parameter α, which characterizes the leaky wave supported in the grating region. The determination of α has been discussed in the preceding report. The present report focuses on the derivation of simplified but accurate formulae for α and on the design criteria that can thereby be inferred for fabricating gratings having desirable characteristics.

Usually, the need for constructing a grating coupler occurs when a planar thin-film optical guide is given, i.e., the equivalent refractive index $N = \beta_{sw}/k_o$ of that guide is known. The question that arises in that case is to design a grating with periodicity d and height t_g, so that d and t_g must be determined in such a manner as to reach a desired value for the leakage parameter α. In addition, one usually desires to have a single diffraction order in order to maximize the coupling efficiency; this is achieved by requiring that

$$|N - (\lambda/d)| < \sqrt{\varepsilon_a} = 1 \quad , \tag{1a}$$

and

$$2(\lambda/d) - N > \sqrt{\varepsilon_s} > 1 \quad , \tag{1b}$$

where the notation is identical to that in the preceding report.

Conditions (1) are summarized in Fig. 1, wherein the shaded regions indicate the range of values for N and λ/d that lead to a single outgoing beam in each of the two air regions above and below the guide-grating structure on which a surface wave is incident, i.e., in an output coupler. By reciprocity, a beam incident on an input coupler will satisfy analogous conditions for maximizing the coupling performance.

After establishing a suitable value of d/λ via Fig. 1, it now remains to determine a corresponding value of t_g/λ that would yield the desired leakage parameter α. For this purpose, we first consider gratings having rectangular profiles. Using Eq. (30) from the preceding report, we then find:

(1) <u>For $a_g t_g \tilde{<} 1$:</u>

$$\alpha_t^{(c)}{}_\lambda = \begin{cases} \dfrac{2(\varepsilon_f - N^2)(\varepsilon_r - \varepsilon_a)^2}{N(\varepsilon_f - \varepsilon_a)} \dfrac{\Lambda}{t_{eff}} \left(\dfrac{t_g}{\lambda}\right)^2 , & \text{for TE modes;} \\ \dfrac{2(\varepsilon_f - N^2)(\varepsilon_r - \varepsilon_a)^2}{N\varepsilon_f\varepsilon_g} \dfrac{N^2 - \varepsilon_a + (N\Lambda/\lambda)^2[N - (\lambda/d)]^2}{N^2 - \varepsilon_a + (\varepsilon_a/\varepsilon_f)^2(\varepsilon_f - N^2)} \dfrac{t_g^2}{\Lambda t_{eff}} , & \text{for TM modes.} \end{cases} \tag{2}$$

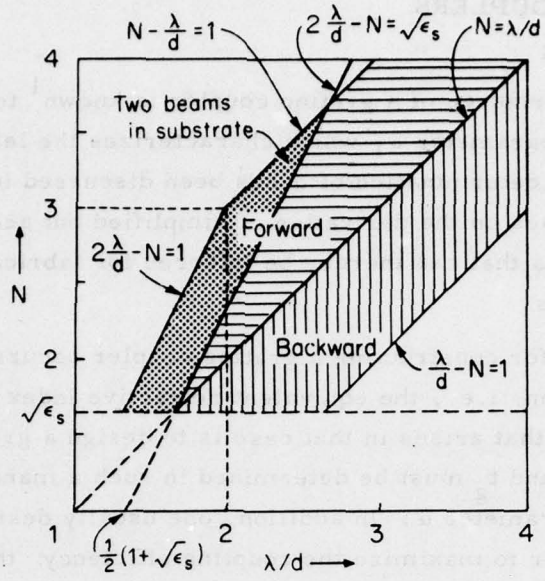


Fig. 1. $N = \beta_{sw}/k_o$ versus λ/d, indicating range of values (shown shaded) for exciting a single diffraction order above the grating. Forward and backward leakage conditions are indicated by different shadings.

Here the superscript c in $\alpha_t^{(c)}$ indicates that the result applies to the canonic rectangular profile having $d_1 = d/2$, whereas the subscript t indicates that Eq. (2) holds for small values of t_g.

(2) For $a_g t_g \gtrsim 1$:

$$\alpha_T^{(c)}\lambda = \begin{cases} \dfrac{\epsilon_f - N^2}{2\pi^2 N(\epsilon_f - \epsilon_g)} \dfrac{(\epsilon_r - \epsilon_a)^2}{2N - (\lambda/d)} \dfrac{d}{\lambda t_{eff}} & \text{for TE modes;} \\ \dfrac{\epsilon_f - N^2}{2\pi^2 N\epsilon_f\epsilon_g} \dfrac{(\epsilon_r - \epsilon_a)^2}{2N - (\lambda/d)} \dfrac{N^2 - \epsilon_g + (N\Lambda/\lambda)^2[N - (\lambda/d)]^2}{N^2 - \epsilon_g/\epsilon_f)^2(\epsilon_f - N^2)} \dfrac{\lambda d}{\Lambda t_{eff}}, & \text{for TM modes.} \end{cases} \tag{3}$$

Again, the superscript c in $\alpha_T^{(c)}$ refers to the canonic rectangular profile with $d_1 = d/2$; the subscript T indicates that Eq. (3) holds for large t_g.

Relations (2) and (3) are basic and contain grating parameters, all of which are readily known except for t_{eff}. The value of t_{eff} may be (somewhat grossly) approximated by t_f; for a better approximation, we use Eq. (11) with m = a in Eq. (2), whereas in Eq. in Eq. (3) we utilize Eq. (11) with m = g. Here Eq. (11) refers to that given for t_{eff} in the preceding report.

The accuracy of the above approximations is illustrated in Fig. 2, which describes results for a typical grating operating in the fundamental TE_0 mode. Leakage occurs only in the n = -1 order. The solid curves have been obtained by using the perturbation procedure presented in the preceding report, whereas the points denoted by X signs denote values calculated by a rigorous approach.[2] The approximations (2) and (3) are shown by dashed curves. The upper and lower bounds $\overline{\alpha}\lambda$ and $\underline{\alpha}\lambda$, respectively, refer to limit values that can easily be determined via the parameter B_n of Eq. (30) in the preceding report.

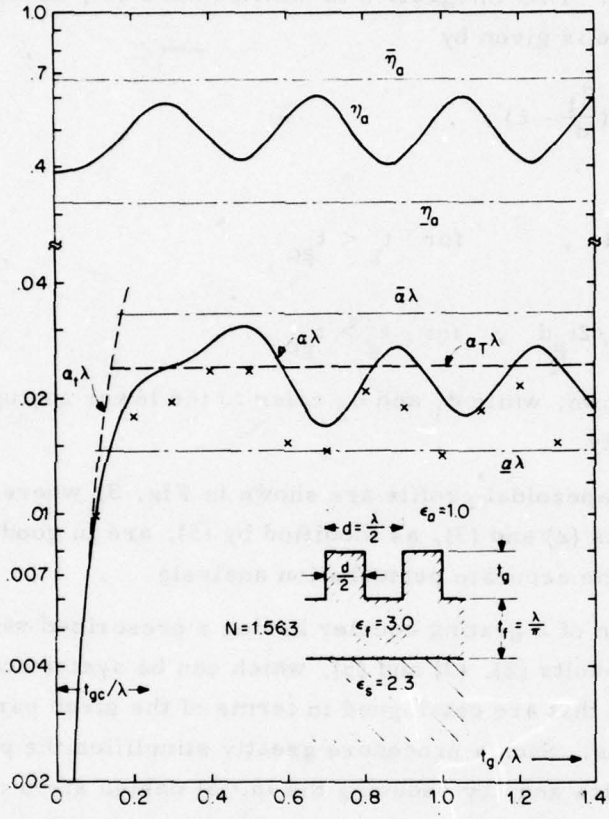


Fig. 2. Variation of normalized leakage $\alpha\lambda$ and power ratio η_a versus t_g λ, for the TE_0 mode in a canonic grating.

Figure 2 describes a typical result for TE modes and similar results hold for TM modes. It is also noted that the efficiency η_a of coupling into the upper air region varies around 50%, as predicted at the end of the preceding report. It is significant to observe that the dashed lines representing Eqs. (2) and (3) in Fig. 2 are very close to the solid line representing the more accurate perturbation result. Furthermore, the latter results are less than 30% higher in value than the exact results shown by the X

signs in Figure 2. Such an accuracy is well within that required for design purposes and is consistent with achievable tolerances.[1]

For gratings having profiles that are different from rectangular, results (2) and (3) can be extended provided the different profile is symmetrical. In that case, the new leakage factor α is given in terms of $\alpha^{(c)}$ of the rectangular profile by

$$\alpha = \rho^2 \alpha^{(c)} \quad , \tag{4}$$

where ρ is a reduction factor which can be readily calculated via an integration of $\varepsilon(x, z)$ over the profile shape. This integration is omitted here but, as an example, the result for a trapezoidal shape is given by

$$\rho = \frac{\sin \pi\delta}{\pi\delta} \sin \pi\left(\frac{d_1}{d} - \delta\right) \quad , \tag{5}$$

where

$$\delta = \begin{cases} (d_1 - d_2)/2d \quad , & \text{for} \quad t_g < t_{gc} \\ (d_1 - d_2)t_{gc}/2t_g d \quad , & \text{for} \quad t_g > t_{gc} \end{cases} \quad . \tag{6}$$

Here $t_{gc} = 1/a_g$ is known, while d_1 and d_2 refer to the lower and upper basis of the trapezoid, *respectively*.

Results for a trapezoidal profile are shown in Fig. 3, where again it is seen that the dashed-line results (2) and (3), as modified by (5), are in good agreement with the solid-line results of the accurate perturbation analysis.

The construction of a grating coupler having a prescribed value of α can thus be based on the use of results (2), (3) and (5), which can be systematized into a series of design data or graphs that are catalogued in terms of the given parameters N, ε_f, ε_s and suitable profile shapes. Such a procedure greatly simplifies the prediction of the coupling characteristics and, by reducing the initial design steps to a set of simple calculations of algebraic relationships, it avoids the use of complex computer programs and very time consuming calculations by means of high-precision computers.[2]

More elaborate details of the aspects presented here are given in a comprehensive paper, where the effects of asymmetric (blazed) grating profiles are also discussed.[3]

National Science Foundation
ENG74-23908

Office of Naval Research
N00014-75-C-0421

T. Tamir

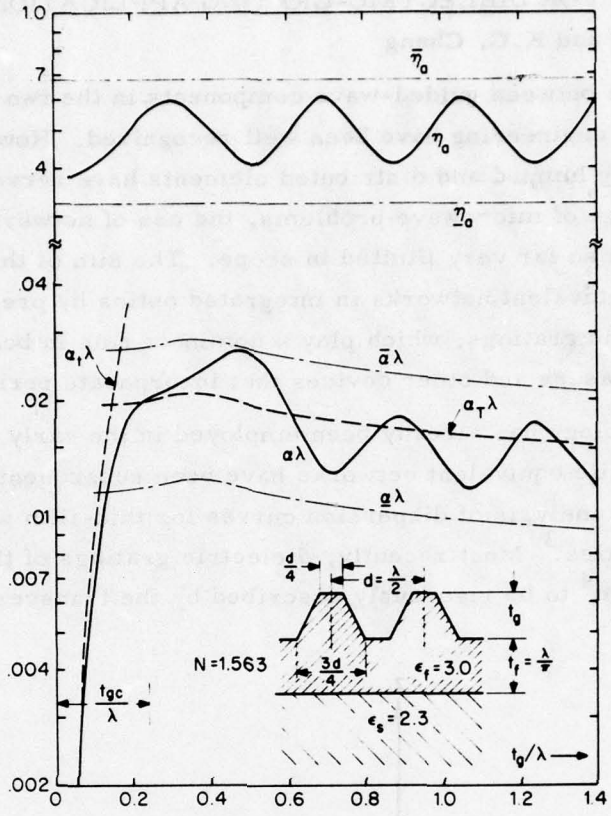


Fig. 3. Variation of normalized leakage $\alpha\lambda$ and power ratio η_a versus t_g λ for the TE_0 mode in a grating having a symmetric trapezoidal profile.

REFERENCES

1. For background and an extensive bibliography up to early 1975, see T. Tamir, Ed., "Integrated Optics," (New York: Springer-Verlag, 1975) Chap. 3, p. 84.

2. S.T. Peng and T. Tamir, "Theory of Periodic Dielectric Waveguides," IEEE Trans. Microwave Theory Tech., Vol. MTT-23, pp. 123-133 (January 1975).

3. T. Tamir and S.T. Peng, "Analysis and Design of Grating Couplers," Appl. Phys., Vol. 14, pp. 235-254 (November 1977).

NETWORK METHODS FOR DIELECTRIC-GRATING APPLICATIONS

T. Tamir, S. T. Peng and K. C. Chang

The similarities between guided-wave components in the two areas of integrated optics and microwave engineering have been well recognized. However, while equivalent networks involving lumped and distributed elements have served as a powerful tool in treating a wide range of microwave problems, the use of network methods in integrated optics has been so far very limited in scope. The aim of this report is to show the effectiveness of equivalent networks in integrated optics by presenting their novel application to dielectric gratings, which play a dominant role in beam couplers, filters, distributed-feedback lasers and other devices that incorporate periodic structures.

Network terminology has already been employed in the early microphotolithographic work at infra-red,[1] while equivalent networks have been subsequently used mostly in transverse-resonance analysis of dispersion curves for thin-film waveguides of the strip[2] or planar varieties.[3] Most recently, dielectric gratings of the type sketched in Fig. 1 have been shown[4] to be rigorously described by the transverse network configu-

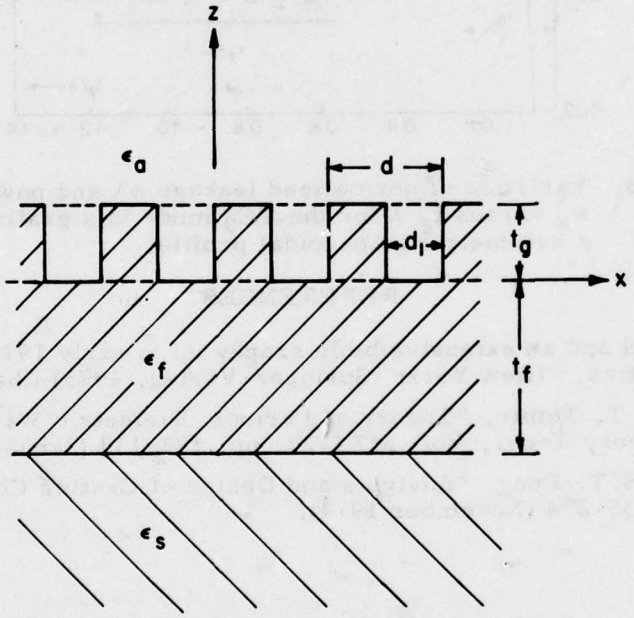


Fig. 1. Typical dielectric gratings used in applications for integrated optics.

ration given in Figure 2. Here B denotes a lossless network that couples all of the space harmonics, each of which is represented by a transmission-line circuit in the regions outside the grating. Whereas the complete network is rather complicated if a

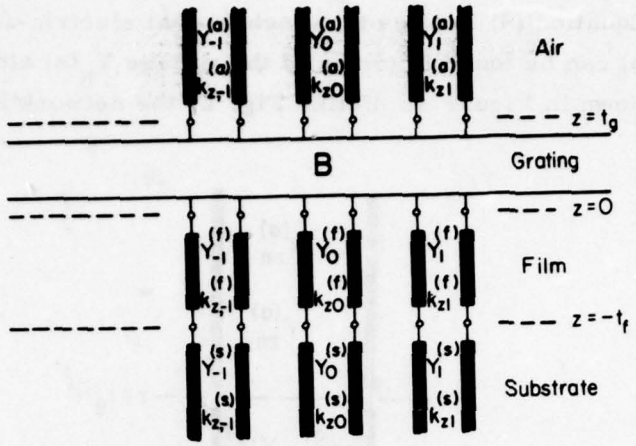


Fig. 2. Equivalent network representation of the dielectric-grating configuration.

large number of space harmonics must be accounted for, it nevertheless leads to systematic computational procedures for structures with a larger number of layers and/or for gratings having more complex profiles.

A great simplification in the equivalent network of Fig. 2 can be achieved by adopting a perturbation approach[5] which assumes that the grating region appears as a modification of a uniform layer having a relative permittivity ε_g. One can then express the permittivity in that layer as

$$\varepsilon(x) = \varepsilon_g\left[1 + \sum_{n\neq 0} \varepsilon_n \exp\left(i\,\frac{2n\pi}{d}\,x\right)\right] \quad , \tag{1}$$

where the summation is regarded as a perturbation term. The electric field $E = E_y$ for TE modes can be written as

$$E(x, z) = \sum_n V_n(z)\, \exp(i\,\kappa_n x) \quad , \tag{2}$$

where

$$\kappa_n = \kappa_0 + \frac{2n\pi}{d} \quad , \quad n = 0, \pm 1, \pm 2, \cdots \quad , \tag{3}$$

and κ_0 is generally complex.

If Eqs. (1) and (2) are introduced into the wave equation and only first-order terms are retained, one gets

$$\frac{d^2 V_n(z)}{dz} + (k^2 - \kappa_n^2)\, V_n(z) = -k^2\, \varepsilon_n\, V_0(z) \quad , \tag{4}$$

which holds for all $n \neq 0$. However, $V_0(z)$ represents a presumably known (unperturbed) incident field. Equation (4) implies that each partial electric-field (space harmonics) $E_n(x, z)$ in Eq. (6) can be found in terms of the voltage $V_n(z)$ along the equivalent transverse network shown in Figure 3. Unlike Fig. 2, the network in Fig. 3 is a simple,

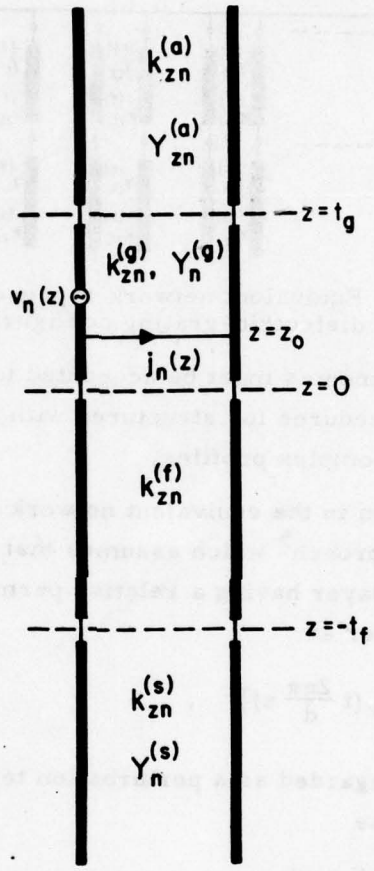


Fig. 3. Simplified network representation for TM-mode problems. For TE modes, the network is unchanged, but $v_n(z) = 0$.

uncoupled and independent transmission-line circuit characterized by propagation factors $k_{zn}^{(u)}$ and characteristic admittances $Y_n^{(u)} = k_{zn}^{(u)}/\omega \mu_o$, where

$$k_{zn}^{(u)} = (k^2 \varepsilon_u - \kappa_n^2)^{1/2} \quad . \tag{5}$$

Here the index u = a, g, f or s denotes the air, grating, film or substrate region, respectively. The voltage $V_n(z)$ is determined by the distributed current $j_n(z)$, which is proportional to the right-hand side of Eq. (4) and is known for any given incident voltage

$V_0(z)$. A similar treatment holds for TM-modes, except that the voltage $V_n(z)$ is then determined by both distributed currents $j_n(z)$ and voltages $v_n(z)$ inside the grating region ($0 < z < t_g$), as indicated in Figure 3.

The principal advantage of this simplified approach is that it can be easily applied to any given excitation problem. Thus, the incident field may be a plane wave impinging obliquely on the grating (for scattering phenomena), a surface wave traveling longitudinally (for beam-coupling or filtering applications), or a standing wave along the grating (for distributed-feedback lasers). Each one of these situations is distinguished from the other only by the fact that the voltage and current sources are different. However, these sources are known in every case, and they are simply prescribed by the given incident field. Because of their relative simplicity, equivalent networks such as that in Fig. 3 are extremely helpful in providing both physical insight and accurate quantitative evaluations for practical design problems.

As an example of an optical beam-coupler application, which requires a leaky-wave regime, Fig. 4 shows the normalized attenuation $\alpha d/2\pi$ of leaky waves supported

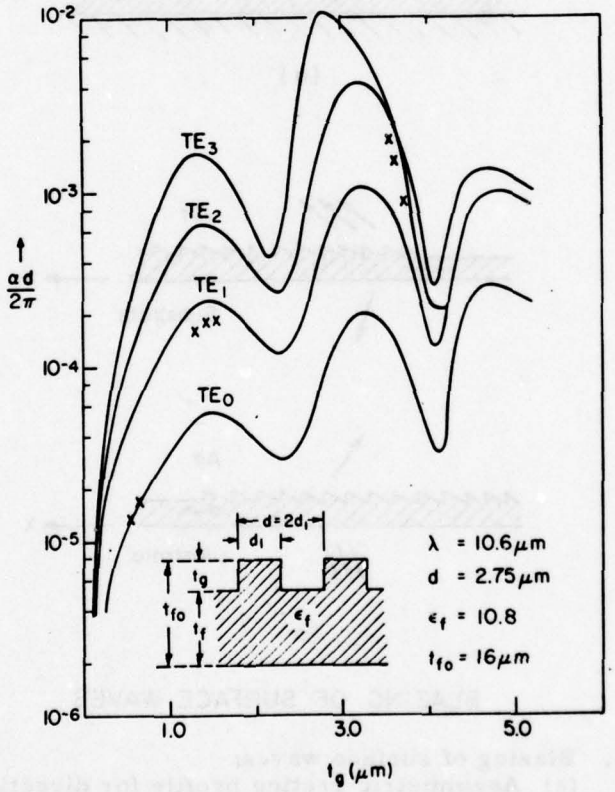


Fig. 4. Normalized leakage $\alpha d/2\pi$ versus grating thickness t_g for several modes guided by a corrugated GaAs slab. Points X indicate results obtained by the exact analysis.

by a corrugated GaAs wafer used in converting a light beam into a surface wave, or vice-versa. In this case, the minima and maxima of the curves can be directly related to the height t_g of the grating. As t_g varies, the transmission line between $z = 0$ and $z = t_g$ in Fig. 3 also varies; hence the leaked energy (and therefore α) exhibits maxima and minima because of the cyclic variations of the input impedances "seen" by the sources.

Another important advantage of this network approach is that it can handle gratings with arbitrary periodic profiles. Consequently, results can be obtained for configurations that, unlike the rectangularly corrugated gratings of Fig. 1, do not lend themselves to an exact solution of the pertinent boundary-value problem.

A particularly interesting case is the asymmetrical triangular profile shown in Fig. 5(a), which can produce directional discrimination ("blazing") of incident surface

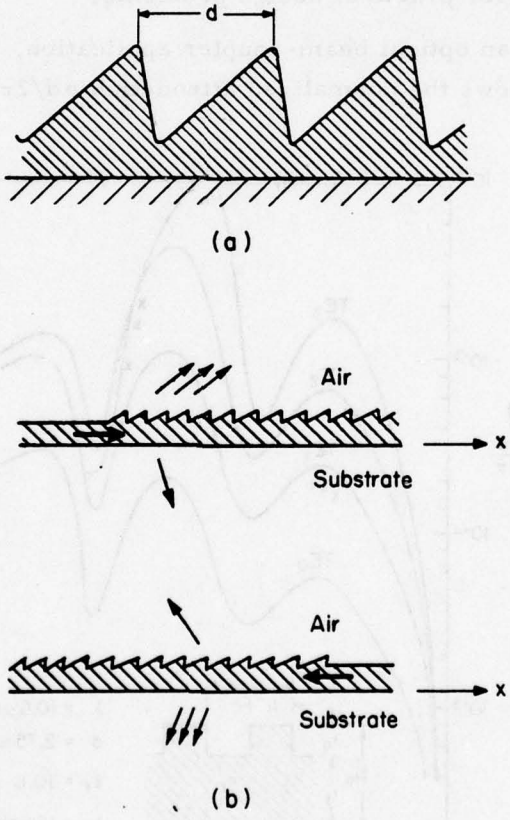


Fig. 5. Blazing of surface waves:
(a) Asymmetric grating profile for directional discrimination of incident surface waves;
(b) Result of grating asymmetry, showing that a surface wave incident from the right is leaked mostly into substrate whereas a surface wave incident from the left is leaked mostly into the air region.

waves. This is achieved by designing the grating so as to induce a stronger energy leakage into either the upper (air) or the lower (substrate) regions, depending on whether incidence is from the right or left, respectively, as suggested in Figure 5(b). In network terms, the desired blazing effect is obtained by choosing the sources $j_n(z)$ and $v_n(z)$ so as to direct the power flow selectively into either the air or the substrate region. For example, the sources can be judiciously distributed along the transmission line segment $0 < z < t_g$ so as to interfere destructively for the field in the substrate region while interfering constructively for the field in the air region. After determining a suitable source distribution, the geometrical shape of the grating profile can be obtained by means of Fourier transforms.

To summarize, the greatest advantage of the equivalent-network representation is that it permits a direct interpretation of the grating characteristics in terms of the constituent parts of the transmission-line circuit. Thus, the leakage properties of grating couplers, the asymmetric behavior of triangular gratings, the Bragg-reflection regime of distributed-feedback lasers, etc. can all be phrased in terms of the transmission-line parameters and the fields associated with them. By extension, the application of such network methods to a wider class of integrated-optics configurations is expected to provide a powerful analytic technique for design-oriented purposes.

Joint Services Technical Advisory Committee
F44620-74-C-0056

T. Tamir, S.T. Peng and K.C. Chang

National Science Foundation
ENG74-23908

Office of Naval Research
N00014-75-C-0421

REFERENCES

1. D.B. Anderson, "Applications of Microphotolithography to Millimeter and Infrared Devices," Proc. IEEE, Vol. 54, pp. 657-661 (April 1966).
2. E.A.J. Marcatili, "Dielectric Rectangular Waveguide and Directional Coupler for Integrated Optics," Bell Syst. Tech. J., Vol. 48, pp. 2071-2102 (September 1969).
3. T. Tamir, "Inhomogeneous Wave Types at Planar Interfaces: II - Surface Waves," Optik, Vol. 37, pp. 204-228 (February 1973).
4. S.T. Peng. T. Tamir and H.L. Bertoni, "Theory of Periodic Dielectric Waveguides," IEEE Trans. Microwave Theory Tech., Vol. MTT-23, pp. 123-133 (January 1975).
5. S.T. Peng and T. Tamir, "TM-Mode Perturbation Analysis of Dielectric Gratings," Appl. Phys., Vol. 6, pp. 35-38 (May 1975).

ON SIMULTANEOUS BLAZING OF TRIANGULAR GROOVE DIFFRACTION GRATINGS

L. S. Cheo, J. Shmoys and A. Hessel

Numerical results for plane wave scattering from a perfectly conducting diffraction grating with triangular groove profile show that simultaneous blazing with 100% efficiency in both polarizations is obtainable in a Littrow arrangement. The range of angles of incidence, for which such simultaneous blazing occurs, is considerably narrower than that obtainable with a rectangular groove profile.

A. Introduction

In a previous paper[1] it was shown that a rectangular-groove perfectly conducting grating can be blazed to diffract all of the incident energy into the n = -1 spectral order, provided two conditions are met. One of these is the Bragg condition, equivalent to placing the grating in a Littrow mount; the other is that the depth of the grating, h, be adjusted to the correct value, which is a function of the angle of incidence θ, and of the profile. Furthermore it was pointed out that by examining the curves h vs. θ for both TE(P) and TM(S) polarizations and locating their intersections, conditions for simultaneous perfect blazing could be found. Further calculations of simultaneous perfect blazing by a rectangular groove grating together with its experimental verification in the microwave region was recently published by Jull et al.;[2] additional calculated performance data, including the effect of imperfect conductivity was provided by Roumiguieres et al.[3]

The question of possible simultaneous blazing of triangular groove gratings was raised by Stroke[4] who indicated that somewhat shallower gratings than a standard echelette give better overall results even though they do not have the perfect blazing property in the TM polarization. Echelette (right angle triangular) grating efficiency was also studied theoretically by Maystre and Petit[5] whose results indicate that simultaneous blazing does not occur in this case. McPhedran and Waterworth[6] examined theoretically the efficiency of triangular profile gratings and came to the conclusion that for good performance in the TE polarization a grating deeper than the echelette is desired. They did not search for conditions for perfect blazing.

In the present work dealing with a perfect blazing of perfectly conducting, infinite triangular gratings, we have followed the conjecture of our previous paper[1] that under the constraint of the Bragg condition $\lambda = 2d \sin\theta$ (i.e. of Littrow mount) a grating of any groove shape can be blazed perfectly simply by varying the groove depth h. In the present context this implies keeping ℓ/d constant (cf. Fig. 1), setting λ to $2d \sin\theta$, and varying h to find a zero of the specular reflection coefficient. It may be worth a

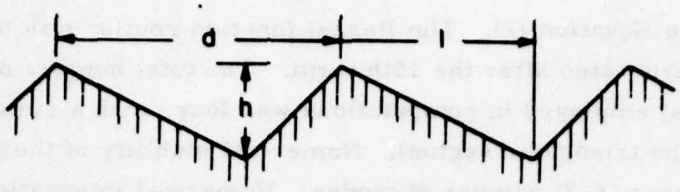


Fig. 1 Triangular-groove grating.

comment that a right angle echelette, because of the constraint imposed on its depth would not be expected to yield simultaneous blazing in both polarizations. This is in fact observed in Fig. 2, which shows the TE efficiency of right angle echelette gratings 100% blazed in the TM polarization as a function of groove depth. The data in this figure has been taken from Reference 5.

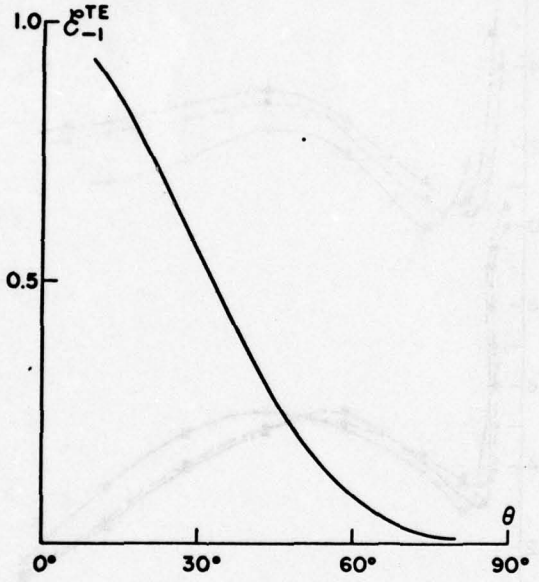


Fig. 2 Efficiency in TE(P) polarization vs. blaze angle of an echelette grating blazed for TM(S) polarization.

B. Numerical Results

The numerical results of this report were obtained using the theoretical development of Jovićević and Sesnic[7] with minor modifications in formulating the set of simultaneous Equations (12) and (13), (32) and (33). To accommodate the nonrectangular triangle groove shape, a Bessel junction J_ν routine for nonintegral values of index was developed, thus extending the applicability and efficiency of the computer

algorithm based on Equation (7). The Bessel function routine was based on the power series expansion truncated after the 15th term. The total number of space harmonics (diffraction orders) employed in computations was four, with a similar number of wedge modes (in the triangular region). Numerical stability of the solution was spot-checked with a larger (6-8) number of modes. Numerical integration of scalar products involved 26 mesh points over the grating period. Computations were performed on a system IBM 370/158, in a time sharing mode, using PL/I. The average CPU time per point was 6sec.

Figure 3 shows the required depth h/d for 100% blazing in the TM(S) polarization



Fig. 3 Relative depth, h/d, vs. angle of incidnece, θ, of a triangular groove grating 100% blazed for TM(S) polarization. Bragg condition, $\lambda/2d = \sin\theta$ was maintained; $\ell/d = 0$ ✕—✕, $\ell/d = 0.25$ ●—●, and $\ell/d = 0.5$ ▲—▲.

for $\ell/d = 0, .25$ and $.5$. We see here two disjoint branches. In Refs. 1-3 only the lower branch was shown for rectangular groove profile. Actually an infinity of higher branches is expected, as can be clearly seen on the rectangular groove grating example. We observe that the character of the curves in both branches is similar;

furthermore the lower branch curve is similar to that for the half open rectangular profile, in that with decreasing values of θ the curve h/d rises, goes through a maximum, dips and rises sharply near θ = 20° i.e., near a Rayleigh angle for n=+1 and n=-2 spectral orders. It is of interest to note also that there is a weak dependence on ℓ/d. Figure 4 shows the relative depth behavior for TE modes. Again we see a

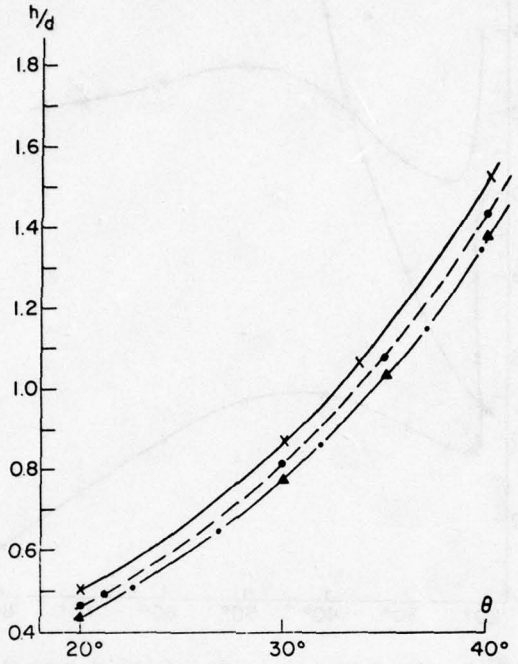


Fig. 4 Relative depth, h/d, vs. angle of incidence θ of a triangular groove grating 100% blazed for TE(P) polarization. Bragg condition, λ/2d = sin θ, was maintained; ℓ/d = 0 ——— , ℓ/d = 0.25 —●— —●— , and ℓ/d = 0.5 —▲— —▲— .

similarity to the half open rectangular grating in that the rise is monotonic, and that the effect of ℓ/d is weak. Figure 5 is a superposition of both the TM and TE blaze curves for ℓ/d = .25. It clearly shows that intersections exist, i.e., simultaneous blazing is feasible but that the physically meaningful operating condition would probably correspond to the intersection of the upper TM branch with the TE curve. The lower branch simultaneous blaze point on the steep part of TM curve would be expected to be quite susceptible to e.g. edge effects, loss and grating imperfections. Numerical data for the simultaneous blaze points found is shown in Table I.

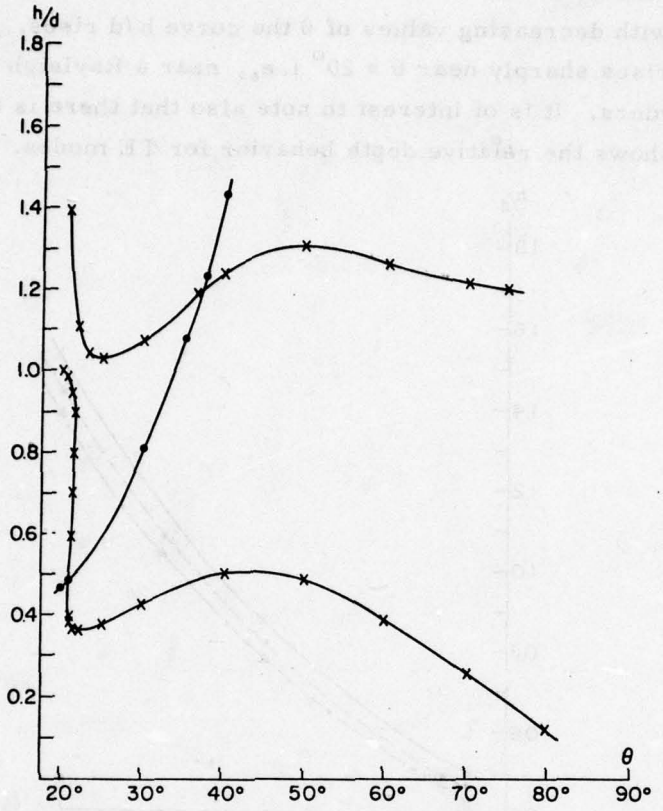


Fig. 5 Relative depth, h/d vs. angle of incidence θ of a 100% blazed triangular grating, under Bragg condition $\lambda/2d = \sin\theta$ and with $\ell/d = 0.25$, for both polarizations: TE(P) —o—o— and TM(S) —X—X— .

TABLE I. Parameters of simultaneously blazed triangular groove gratings.

	Vertex location ℓ/d	Normalized depth h/d	Angle of incidence θ
Lower branch	0.0	0.55	21.5°
	0.25	0.495	21.17°
	0.5	0.46	20.5°
Upper branch	0.0	1.065	33.48°
	0.25	1.195	36.75°
	0.5	1.238	38.1°

C. Conclusions

The conjecture of Ref. 1 with respect to simultaneous blazing of gratings with an arbitrary groove profile has been verified numerically for the case of a triangular shape. It has been shown that simultaneous 100% blazing of a perfectly conducting infinite grating with a triangular profile is possible. However, unlike that for the rectangular shape, which permits a simultaneous blazing for a wide range of incidence angles, such angular range is more restricted, for a triangular groove profile, because of the weak dependence on ℓ/d.

Joint Services Technical Advisory Committee
F44620-74-C-0056 L.S. Cheo, J. Shmoys and A. Hessel
Faculty Research Grant, William Patterson College

REFERENCES

1. A. Hessel, J. Shmoys, and D.Y. Tseng, "Bragg-Angle Blazing of Diffraction Gratings," J. Opt. Soc. Am., Vol. 65, pp. 380-384 (1975).
2. E.V. Jull, J.W. Heath and G.R. Ebbeson, "Gratings that Diffract All Incident Energy," J. Opt. Soc. Am., Vol. 67, pp. 557-560 (1977).
3. J.L. Roumiguieres, D. Maystre and R. Petit, "On the Efficiencies of Rectangular-Groove Gratings," J. Opt. Soc. Am., Vol. 66, pp. 772-775 (1976).
4. G.W. Stroke, "Attainment of High Efficiencies in Blazed Optical Gratings by Avoiding Polarization in the Diffracted Light," Phys. Lett., Vol. 5, pp. 45-48 (1963).
5. D. Maystre and R. Petit, "Étude Quantitative de L'efficacité du Réseau Échelette dans un Montage a Deviation Constante," Nouv. Rev. d'Opt. Appl., Vol. 2, pp. 115-120 (1971).
6. R.C. McPhedran and M.D. Waterworth, "Blaze Optimization for Triangular Profile Grating," Optica Acta, Vol. 20, pp. 177-191 (1973).
7. S. Jovićević and D. Sesnic, "Diffraction of a Parallel and Perpendicular-Polarized Wave from an Echelette Grating," J. Opt. Soc. Am., Vol. 62, pp. 865-877 (1972).

RADIATION PATTERNS OF DIELECTRIC SLAB WAVEGUIDES

L. Bergstein

A. Introduction

In an earlier report[1] we have developed a method which allows the determination of the radiation patterns of terminated dielectric slab waveguides. As an example of the application of the method,[2] we have determined the radiation patterns of two waveguides, one of optical thickness $nd = 0.20\lambda$ and one of thickness $nd = 1.20\lambda$. The results show that the radiation patterns of the waveguide of thickness 0.20λ are considerably wider than those of the waveguide of thickness 1.20λ, although exactly the opposite is to be expected; as the width of a waveguide increases its radiation pattern decreases. In view of the importance of this problem in integrated optics, we have decided to investigate this question in more detail.

A close examination of the problem at hand shows that for dielectric waveguides of practical interest in the optical region, the power carried by the non-radiating modes and in the waves reflected at the waveguide termination is small compared to the power carried in the propagating modes. For the approximate determination of the radiation patterns, we can therefore neglect these waves and consider only the radiating modes and the evanescent waves coupled to them. The error in the results so obtained should be in the order of a few percent. Indeed, this is verified by a comparison of our results with those obtained by more exact but highly cumbersome methods.

B. Determination of the Radiation Patterns

Consider a dielectric slab waveguide of thickness d and refractive index n immersed in a medium of refractive index $n_1 < n$. The waveguide is terminated and radiates into a region of refractive index n_2. As shown in Fig. 1, we choose a Cartesian

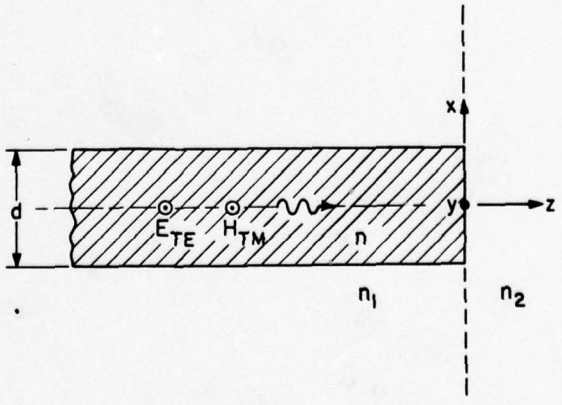


Fig. 1. Geometry of the dielectric slab waveguide.

coordinate system with the z-axis coinciding with the axis of the waveguide and assume that there is no field variation in the y direction. Such a waveguide can support non-coupled TE and TM modes, with

$$\underset{\sim}{E} = [0,\ E_y,\ 0] \quad \text{and} \quad \underset{\sim}{H} = [H_x,\ 0,\ H_z] \tag{1a}$$

for the TE modes, and

$$\underset{\sim}{E} = [E_x,\ 0,\ E_z] \quad \text{and} \quad \underset{\sim}{H} = [0,\ H_y,\ 0]$$

for the TM modes. If we let $\psi(x, z)$ to be equal to either E_y (for TE modes) or to H_y (for TM modes), it is readily found upon application of Maxwell's equation and the appropriate boundary conditions that

$$\psi(x, z) = \begin{cases} (-1)^m \cos\phi\, e^{iku_o z}\, e^{k\xi(x+\frac{d}{2})}, & \text{for } x < -\frac{d}{2}, \\ e^{iku_o z} \cos(kw_o x - m\frac{\pi}{2}), & \text{for } -\frac{d}{2} \le x \le +\frac{d}{2}, \\ \cos\phi\, e^{iku_o z}\, e^{-k\xi(x-\frac{d}{2})} & \text{for } \frac{d}{2} < x. \end{cases} \tag{2}$$

In these equations, λ is the free-space wavelength, $k_o = 2\pi/\lambda$ is the free-space propation constant, $k = nk_o$, θ_o is the synchronous angle of the surface-wave mode with respect to the normal to the waveguide surfaces (i.e., the x direction), $u_o = \sin\theta_o$, $w_o = \cos\theta_o$, $\alpha_1 = n_1/n$ and $\xi = \sqrt{u_o^2 - \alpha_1^2}$. For even modes, $m = 0, 2, 4, \ldots$, whereas for odd modes, $m = 1, 3, 5, \ldots$. Moreover,

$$\tan\phi = g_1 \frac{\sqrt{u_o^2 - \alpha_1^2}}{w_o}, \tag{3}$$

where

$$g_1 = \begin{cases} 1, & \text{for TE modes} \\ 1/\alpha_1^2, & \text{for TM modes}. \end{cases} \tag{4}$$

The synchronous angle θ_o is determined by the transverse resonance condition,

$$\pi \frac{nd}{\lambda} \cos\theta_o = \phi + m\frac{\pi}{2}, \tag{5}$$

giving

$$\tan\left(\pi \frac{nd}{\lambda} w_o\right) = \tan\left(\phi + m \frac{\pi}{2}\right) = \begin{cases} \tan \phi\ , & \text{for even modes,} \\ -1/\tan \phi, & \text{for odd modes.} \end{cases} \tag{6}$$

We can now proceed to determine the radiation pattern of the waveguide.

In the region $z \geq 0$, the electromagnetic field can be expressed as a superposition of plane waves propagating at angles φ with respect to the z direction. Thus, we set

$$\psi(x, z) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(u)\, e^{ik_2(ux + wz)}\, du\ , \tag{7}$$

where

$$u = \sin \varphi \quad \text{and} \quad w = \cos \varphi\ , \tag{8}$$

and $k_2 = n_2\, 2\pi/\lambda = n_2 k_o$. To determine the radiation pattern $F(u)$ we apply the boundary condition at $z = 0$ and obtain

$$F(u) = k_2 \int_{-\infty}^{+\infty} A(x)\, e^{-ik_2 xu}\, du\ , \tag{9}$$

where $A(x) = \psi(x, 0)$, with $\psi(x, z)$ as given by Equation (2). This gives,

$$F(u) = \alpha_2 \left\{ \cos\phi \left[\frac{e^{-i\pi DU}}{\xi + iu} + (-1)^m \frac{e^{i\pi DU}}{\xi - iu} \right] + e^{im\frac{\pi}{2}} \left[\frac{\sin\left[\pi D(w_o + u)\right]}{w_o + u} \right. \right.$$
$$\left. \left. + (-1)^m \frac{\sin\left[\pi D(w_o - u)\right]}{w_o - u} \right] \right\}\ , \tag{10}$$

where $\alpha_2 = n_2/n$,

$$U = \frac{n_2}{n} u = \alpha_2 u\ , \tag{11}$$

and

$$D = nd/\lambda \tag{12}$$

is the relative optical thickness of the waveguide.

Setting

$$\psi = \pi DU = \pi D \alpha_2 u\ , \tag{13}$$

it is readily found from Eq. (10) that

$$F_{es}(u) = F_{es}(0)\left[\cos\psi - \frac{1}{\xi\pi D}\,\psi\sin\psi\right]\Big/\left[1 + \frac{w_o^2 - \xi^2}{(w_o\xi\pi D)^2}\,\psi^2 - \frac{1}{(w_o\xi\pi^2 D^2)}\,\psi^4\right], \tag{14a}$$

for even TE modes,

$$F_{ep}(u) = F_{ep}(0)\left[\left(1 + \frac{\psi^2}{(u_o\pi D)^2}\right)\cos\psi - \frac{\alpha_1^2}{u_o^2\xi\pi D}\,\psi\sin\psi\right]\Big/\left[1 + \frac{w_o^2 - \xi^2}{(w_o\xi\pi D)^2}\,\psi^2 - \frac{1}{(w_o\xi\pi^2 D^2)^2}\,\psi^4\right], \tag{14b}$$

for even TM modes,

$$F_{os}(u) = F_{es}(0)\left[\sin\psi + \frac{1}{\xi\pi D}\,\psi\cos\psi\right]\Big/\left[1 + \frac{w_o^2 - \xi^2}{(w_o\xi\pi D)^2}\,\psi^2 - \frac{1}{(w_o\xi\pi^2 D^2)^2}\,\psi^4\right], \tag{14c}$$

for odd TE modes, and

$$F_{op}(u) = F_{ep}(0)\left[\left(1 + \frac{\psi^2}{(u_o\pi D)^2}\right)\sin\psi + \frac{\alpha_1^2}{u_o^2\xi\pi D}\,\psi\cos\psi\right]\Big/\left[1 + \frac{w_o^2 - \xi^2}{(w_o\xi\pi D)^2}\,\psi^2 - \frac{1}{(w_o\xi\pi^2 D^2)^2}\,\psi^4\right], \tag{14d}$$

for odd TM modes, where $F_{es}(0) = 2\alpha_2\sqrt{1 - \alpha_1^2}/w_o\xi$ and $F_{ep}(0) = 2u_o\alpha_2\sqrt{1-\alpha_1^2}/\alpha_1 w_o\xi$.

For the dominant modes ($m = 0$), $A(x)$ is positive for all x. The absolute maximum of $F(u)$ for these modes will therefore occur at $u = 0$. Figure 2 shows the radiation patterns $|F(u)|^2$ of the first-order TM modes for two slab waveguides of optical thickness $nd = 0.20\lambda$ and $nd = 1.20\lambda$, respectively. Both waveguides have the same refractive index ratio $n_1/n = 0.40$ and are assumed to radiate into free space, with $n_2/n = 0.40$. For comparison purposes we show the corresponding radiation patterns for these waveguides as found by Angulo[3] and Hu and Bergstein[1,2] using much more cumbersome and involved methods. The agreement is almost perfect. We observe that the waveguide of thickness 1.20λ has a radiation pattern which is approximately four times wider than the pattern of the 0.20λ thick waveguide.

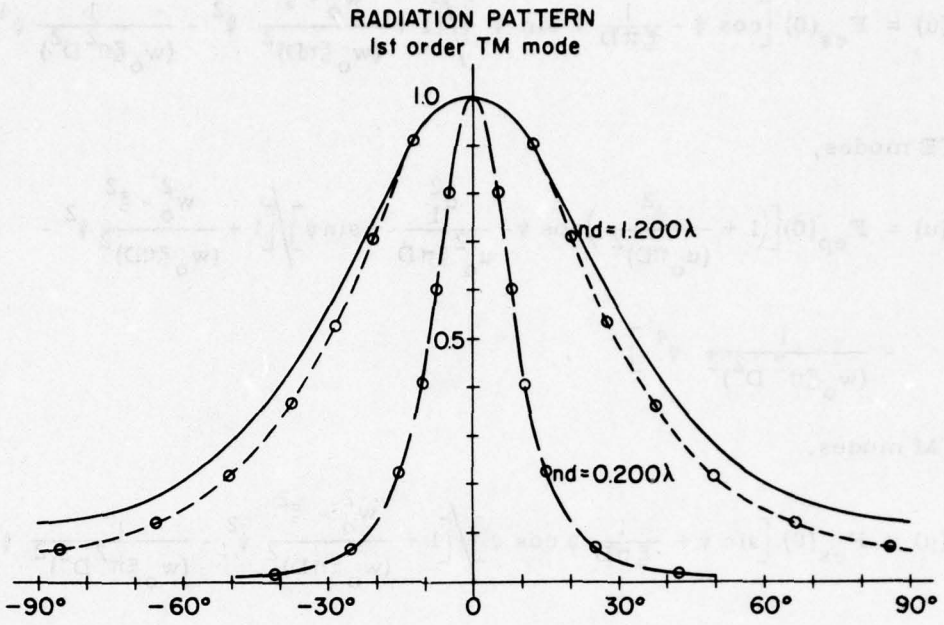


Fig. 2. Radiation patterns $|F(\varphi)|^2$ for the first-order TM modes of a terminated dielectric slab waveguide of refractive index n = 1.578 immersed in a medium of refractive index $n_1 = 1$. The graphs show $|F(\varphi)|^2$ for waveguides of optical thickness $nd = 0.20\lambda$ and 1.20λ. The points marked "0" correspond to the results reported by Angulo and Hu.

In Fig. 3 we show the radiation patterns of the first-order TE modes for waveguides of refractive index ratios $n_1/n \approx n_2/n = 0.40$ and thicknesses $nd_1 = 0.125\lambda$, 0.625λ and 3625λ respectively. Again, it is seen that the radiation pattern of the 0.625λ waveguide is considerably wider than the pattern of the 0.125λ waveguide, and that the 0.125λ and 3.625λ waveguides have approximately the same radiation patterns.

We define the angular beamwidth ABW = $2(\nabla\varphi)$ of the first-order mode radiation patterns as the width of angles for φ for which $|F(u)|^2 \geq \frac{1}{2} F^2(o)$. Using Eqs. (14a) and (14b), the half-power width $\nabla\varphi$ can be readily determined as a function of the waveguide thickness $D = nd/\lambda$. The results are given in Fig. 4, where $2(\nabla\varphi)$ is shown for TE and TM modes as a function of D for various values of the refractive index ratio $\alpha_1 = n_1/n$. It is assumed that the waveguide radiates into free space. For comparison purposes, the figure also shows the half-power width for the limiting case of $n \to \infty$ and $n_1/n \to 0$, that is, the width $2(\nabla\varphi_o)$ of a radiation pattern which would be obtained if we were to neglect the evanescent waves associated with the propagating modes. The field distribution for the first-order modes is then given by

$$\psi_o(x, 0) = \begin{cases} \cos(\pi x/d) , & \text{for } -\frac{d}{2} \leq x \leq +\frac{d}{2} , \\ 0 & \text{otherwise} . \end{cases}$$

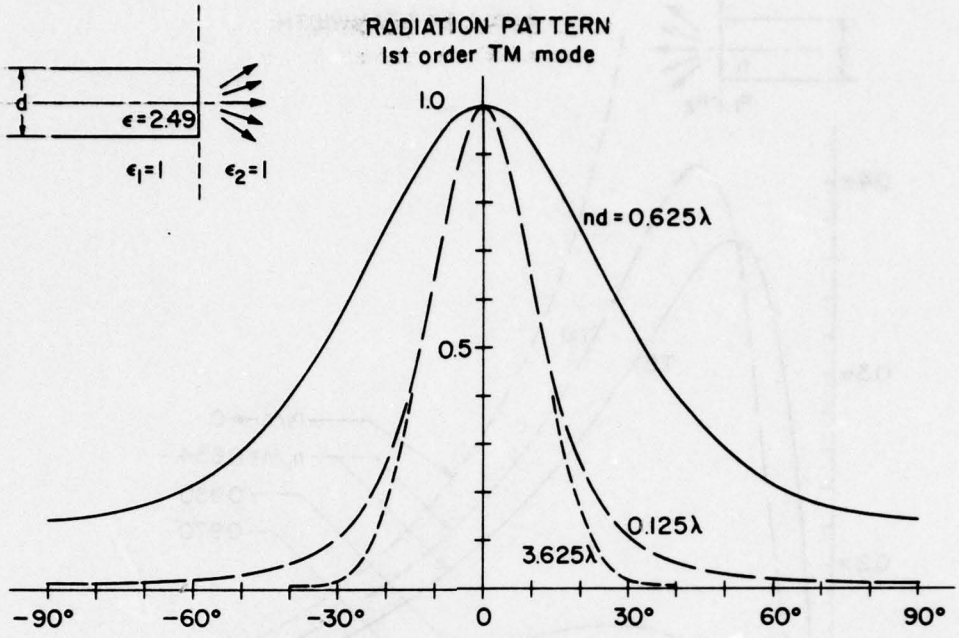


Fig. 3. Radiation patterns $|F(\varphi)|^2$ for the first-order TE modes of a terminated dielectric slab waveguide of refractive index n = 1.578 immersed in a medium of refractive index n_1 = 1. The graphs show $|F(\varphi)|^2$ for waveguides of optical thickness nd = 0.125λ, 0.625λ and 3.625λ.

resulting in a radiation pattern

$$F_o(u) = \cos(\pi Du)/[1 - (2Du)^2] \quad ,$$

and a half-power width

$$2(\nabla\varphi_o) = 2\sin^{-1}(0.594\lambda/n_2 d) = 2\sin^{-1}(0.594/\alpha_2 D) \quad . \tag{15}$$

The results are rather interesting. They show that the width of the radiation pattern first increases with increasing waveguide thickness, reaching a maximum when the optical thickness nd of the waveguide is in the order of a wavelength, and then decreasing as the waveguide thickness further decreases, approaching the value of $2(\nabla\varphi_o)$ as given by Equation (15). As an example, it is seen that when n = 1.578 and n_1 = 1.00, the width of the TM radiation pattern of a waveguide of thickness nd ≅ 0.25λ is the same as that of a waveguide with nd = 4.5λ; the TM pattern width of a waveguide with nd ≈ 0.125λ is the same as that of a slab width nd ≈ 10λ. Corresponding results are obtained for other refractive index ratios. Of course, it is not hard to find a physical explanation for these results.

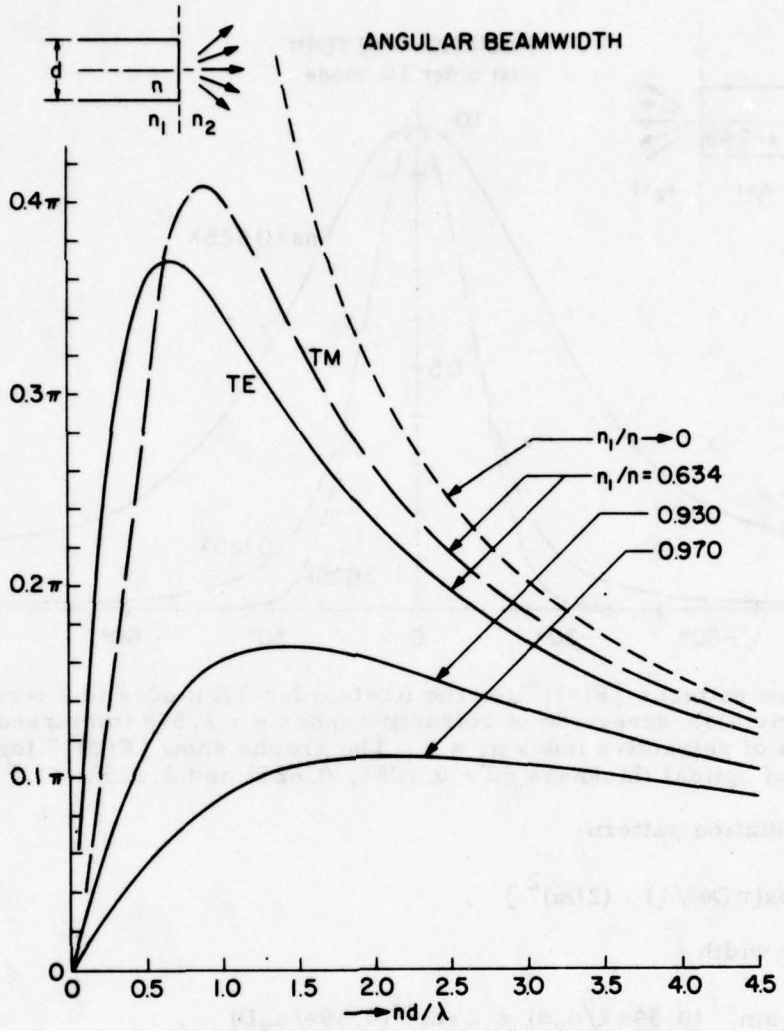


Fig. 4. Angular beamwidth $2(\nabla\varphi)$ of the first-order mode radiation patterns of terminated dielectric slab waveguides as a function of the optical thickness nd/λ of the waveguide. The graphs show $2(\nabla\varphi)$ for various values of the refractive index ratio n_1/n. Also shown in $2(\nabla\varphi_o)$ for the limiting case $n \to \infty$ or $n_1/n \to 0$.

When the waveguide thickness is small, the synchronous angle θ of the surface-wave modes is close to the critical angle $\theta_c = \sin^{-1}(n_1/n)$ of total reflection. As a result, the attenuation factor $\xi = \sqrt{\sin^2\theta - (n_1/n)^2}$ for the evanescent waves outside the waveguide is small and the evanescent "tail" is large, making the waveguide look much wider and the radiation pattern narrower. As the waveguide thickness increases, the synchronous angle increases and so does the attenuation factor ξ. This results in a shorter evanescent tail and in a decrease in the apparent waveguide thickness.

Finally, as the waveguide thickness further increases the effect of the evanescent waves becomes negligible and the width of the radiation pattern decreases with increasing thickness nd.

To further compare our results with those available in the literature, we show in Fig. 5 the radiation patterns (as given by Eq. (14d))of the second-order TM modes

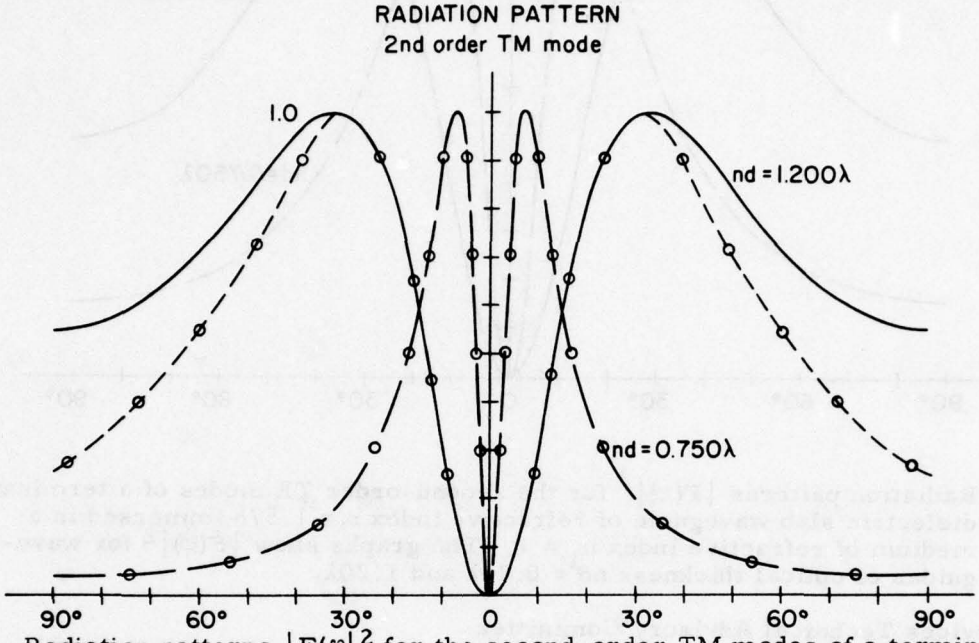


Fig. 5. Radiation patterns $|F(\varphi|^2$ for the second-order TM modes of a terminated dielectric slab waveguide of refractive index n = 1.578 immersed in a medium of refractive index n_1 = 1. The graphs show $|F(\varphi)|^2$ for waveguides of optical thickness nd = 0.75λ and 1.20λ. The points marked "0" correspond to the results reported by Angulo and Hu.

for two slab waveguides of optical thickness nd = 0.75λ and nd = 1.20λ respectively, immersed in and radiating into free space. The results found by Angulo[3] and Hu[2] are also shown in the figure. For the 0.75λ waveguide there is perfect agreement between our results and those of Angulo and Hu, whereas for the 1.20λ waveguide there is a discrepancy in the results at the tail end of the radiation pattern. The radiation patterns of these waveguides for the seond-order TE modes are shown in Fig. 6. We again observe that the TE and TM radiation patterns of the 1.20λ waveguide are considerably wider than the corresponding patterns of the 0.75λ waveguide.

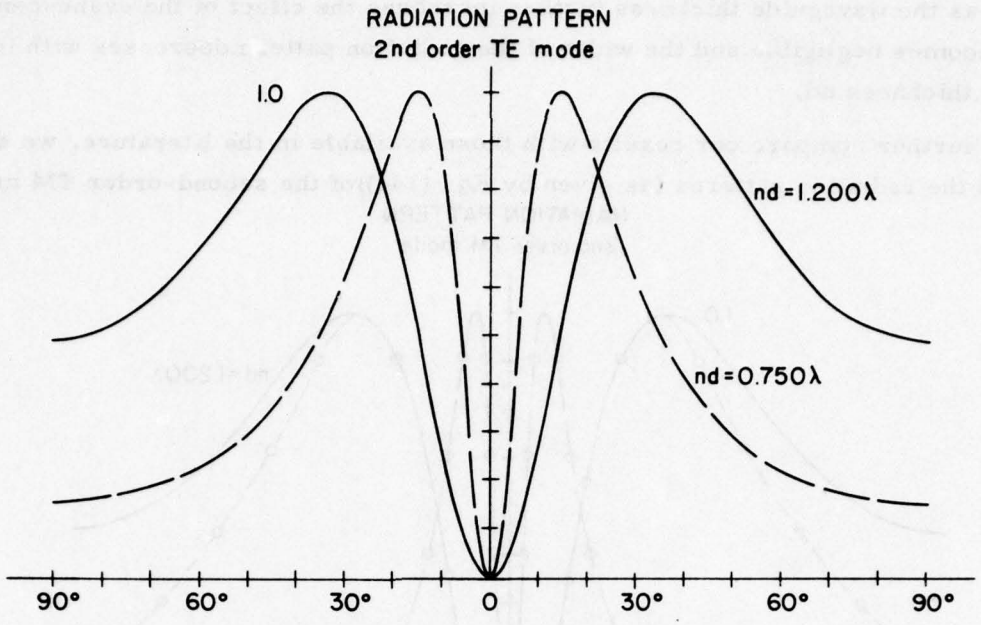


Fig. 6. Radiation patterns $|F(\varphi)|^2$ for the second-order TE modes of a terminated dielectric slab waveguide of refractive index n = 1.578 immersed in a medium of refractive index n_1 = 1. The graphs show $|F(\varphi)|^2$ for waveguides of optical thickness $nd = 0.75\lambda$ and 1.20λ.

Joint Services Technical Advisory Committee
F44620-74-C-0056

L. Bergstein

REFERENCES

1. E-W. Hu and L. Bergstein, "Analysis of Surface Waveguide Discontinuities," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76, pp. 22-26 (1976).

2. E-W. Hu, "Optical Waveguide Discontinuities," Ph.D. Dissertation, Polytech. Inst. of New York (June 1976).

3. C. Angulo, "Diffraction of Surface Waves by a Semi-Infinite Dielectric Slab," Ph.D. Dissertation, Polytech. Inst. of New York (June 1955).

POLARIZATION INDEPENDENT DIELECTRIC CONSTANT PROFILES

L. Bergstein

A. Introduction

An electromagnetic field propagating in linear, homogeneous and isotropic media can be decomposed into two mutually independent constituent fields, one which is transverse electric (TE) and one which is transverse magnetic (TM) with respect to an arbitrarily chosen direction (of propagation).[1] Although the electric and magnetic fields associated with these modes are distinct, the respective intensity distributions are identical and the TE and TM signals propagate with the same group velocity.

A TE-TM decomposition is also possible under certain conditions in inhomogeneous media, as for example when the variation of the dielectric constant and/or the magnetic permeability is one-dimensional. In the inhomogeneous case, however, the intensity distributions of the two fields are usually distinct and the TE and TM signals propagate with different group velocities. We investigate here the question whether separable inhomogeneity profiles can be found for which the TE and TM modes have identical intensity distributions in direction of propagation and propagate with the same group velocity. A waveguide made of such an inhomogeneous medium can thus support a non-polarized signal. We have earlier[2] solved the problem for the case of plane-stratified and rotationally symmetric cylindrical media. We now consider the general case and derive a set of necessary and sufficient conditions for the polarization independent propagation of non-coupled TE and TM modes. As an example, we determine the polarization independent dielectric constant profiles for various geometries.

B. Polarization Independent Propagation

The relation between the electric field $\underset{\sim}{E}$ and magnetic field $\underset{\sim}{H}$ of an electromagnetic wave of radian frequency ω propagating in a charge-free region of a linear and isotropic medium is given by two of Maxwell's equations,

$$\nabla \times \underset{\sim}{E} = i\omega\mu\underset{\sim}{H} \quad , \qquad \nabla \times \underset{\sim}{H} = -i\omega\varepsilon\underset{\sim}{E} \quad , \tag{1}$$

where i is the imaginary unit, ε is the dielectric constant of the medium, μ is its magnetic permeability, and ∇ is the vector differential operator (with $\nabla \times \underset{\sim}{A}$ = curl $\underset{\sim}{A}$, $\nabla \cdot \underset{\sim}{A}$ = div $\underset{\sim}{A}$ and $\nabla\psi$ = grad ψ). The time dependent factor $\exp(-i\omega t)$ is self-understood and is left out.

Maxwell's equations can be simplified when the variation of the dielectric constant ε and/or magnetic permeability μ is only two dimensional. We choose the direction in which there is no variation of the media parameters as the n direction and decompose

the electric and fields into transverse and longitudinal components with respect to that direction. Thus, we set,

$$\underset{\sim}{E} = \underset{\sim}{E}_T + \underset{\sim}{u}_n E_n \quad \text{and} \quad \underset{\sim}{H} = \underset{\sim}{H}_T + \underset{\sim}{u}_n H_n \quad , \tag{2}$$

where $\underset{\sim}{u}_n$ is the unit vector in the n direction, and the subscripts T and n stand for "transverse" and "normal" (i.e., longitudinal) respectively. Moreover, since Maxwell's equations are linear, we can express the field as a superposition of two fields, one which is transverse electric with respect to the n direction, with $E_n = 0$, and one which is transverse magnetic, with $H_n = 0$. It is then readily found that for the transverse electric field, $\underset{\sim}{E}' = \underset{\sim}{E}'_T$,

$$\underset{\sim}{H}'_T = \frac{1}{i\omega\mu}(\nabla_n \times \underset{\sim}{E}') \quad , \quad H'_n = \frac{1}{i\omega\mu}\,\underset{\sim}{u}_n \times \nabla \cdot \underset{\sim}{E}' \quad , \tag{3a}$$

where ∇_n is the n-component of the ∇ operator (for curl operation) and $\underset{\sim}{E}'$ is given by

$$\nabla^2 \underset{\sim}{E}' + \omega^2 \mu\varepsilon\, \underset{\sim}{E}' + \frac{\Delta\mu}{\mu} \times \nabla \times \underset{\sim}{E}' - \nabla(\nabla \cdot \underset{\sim}{E}') = 0 \quad , \tag{4a}$$

with

$$\nabla \cdot (\mu \underset{\sim}{E}') = 0 \quad \text{and} \quad \underset{\sim}{E}' \cdot \left(\frac{\nabla\varepsilon}{\varepsilon} - \frac{\nabla\mu}{\mu}\right) = 0 \tag{5a}$$

Correspondingly, it is found that for the transverse magnetic field, $\underset{\sim}{H}'' = \underset{\sim}{H}''_T$,

$$\underset{\sim}{E}''_T = -\frac{1}{i\omega\varepsilon}(\nabla_n \times \underset{\sim}{H}'') \quad , \quad E''_n = -\frac{1}{i\omega\varepsilon}\,\underset{\sim}{u}_n \times \nabla \cdot \underset{\sim}{H}'' \quad , \tag{3b}$$

$\underset{\sim}{H}''$ being given by

$$\nabla^2 \underset{\sim}{H}'' + \omega^2 \mu\varepsilon\, \underset{\sim}{H}'' + \frac{\nabla\varepsilon}{\varepsilon} \times \nabla \times \underset{\sim}{H}'' - \nabla(\nabla \cdot \underset{\sim}{H}'') = 0 \quad , \tag{4b}$$

with

$$\nabla \cdot (\varepsilon \underset{\sim}{H}'') = 0 \quad \text{and} \quad \underset{\sim}{H} \cdot \left(\frac{\nabla\mu}{\mu} - \frac{\Delta\varepsilon}{\varepsilon}\right) = 0 \quad . \tag{5b}$$

The power flow in the direction of propagation is given by Poynting's vector

$$\underset{\sim}{P} = \underset{\sim}{E}_T \times \underset{\sim}{H}_T \quad , \tag{6}$$

with $\underset{\sim}{E}_T = \underset{\sim}{E}'_T + \underset{\sim}{E}''_T$ and $\underset{\sim}{H}_T = \underset{\sim}{H}'_T + \underset{\sim}{H}''_T$. For media for which Poynting's vector can be expressed as a sum of two terms, one involving only the TE field components and one involving only the TM components, the TE and TM modes are mutually independent. For such "separable" distributions we have, with $\underset{\sim}{P} = \underset{\sim}{P}_{TE} + \underset{\sim}{P}_{TM}$,

$$\underset{\sim}{P}_{TE} = \frac{1}{i\omega\mu}(\underset{\sim}{E}' \times \nabla_n \times \underset{\sim}{E}') \quad \text{and} \quad \underset{\sim}{P}_{TM} = \frac{1}{i\omega\varepsilon}(\underset{\sim}{H}'' \times \nabla_n \times \underset{\sim}{H}'') \quad . \tag{7}$$

Clearly then, in separable distributions the TE and TM signals will have the same intensity distributions and propagate with the same group velocity if, and only if,

$$\frac{H''_T/\sqrt{\epsilon}}{E'_T/\sqrt{\mu}} = \text{constant} \quad , \tag{8}$$

or, equivalently, if $H''_T/\sqrt{\epsilon}$ and $E'_T/\sqrt{\mu}$ satisfy the same wave equation.

We limit our considerations here to dielectric media, when there is no spatial variation in the magnetic permeability μ. We then find that

$$\nabla \cdot \underset{\sim}{E}' = 0 \quad , \quad \underset{\sim}{E}' \cdot \nabla \epsilon = 0 \quad , \tag{9a}$$

$$\nabla \cdot \underset{\sim}{H}'' = 0 \quad , \quad \underset{\sim}{H}'' \cdot \nabla \epsilon = 0 \quad , \tag{9b}$$

and the wave equations for $\underset{\sim}{E}'$ and $\underset{\sim}{H}''$ reduce to

$$\nabla^2 \underset{\sim}{E}' + \omega^2 \mu \epsilon \, \underset{\sim}{E}' = 0 \tag{10a}$$

and

$$\nabla^2 \underset{\sim}{H}'' + \omega^2 \mu \epsilon \, \underset{\sim}{H}'' + \frac{\nabla \epsilon}{\epsilon} \times \nabla \times \underset{\sim}{H}'' = 0 \quad , \tag{10b}$$

respectively. Polarization independence of the signal will now result when the wave equations for $\underset{\sim}{H}''/\sqrt{\epsilon}$ is identical to the wave equation for $\underset{\sim}{E}'$. Rewriting Eq. (10b) in terms of $\underset{\sim}{H}''/\sqrt{\epsilon} \equiv \underset{\sim}{F}$, we obtain

$$\nabla^2 \underset{\sim}{F} + \omega^2 \mu \epsilon \, \underset{\sim}{F} + \frac{1}{2\epsilon} \left[\nabla \epsilon \times \nabla \times \underset{\sim}{F} - \nabla \times \nabla \epsilon \times \underset{\sim}{F} - 1.5 \frac{(\nabla \epsilon)^2}{\epsilon} \underset{\sim}{F} \right] = 0 \quad . \tag{11}$$

Obviously, this equation will reduce to the same form as Eq. (10a) and result in $\underset{\sim}{F} = \underset{\sim}{H}''/\sqrt{\epsilon} = \underset{\sim}{E}'$ if

$$\nabla \epsilon \times \nabla \times \underset{\sim}{F} - \nabla \times \nabla \epsilon \times \underset{\sim}{F} - 1.5 \frac{(\nabla \epsilon)^2}{\epsilon} \underset{\sim}{F} = 0 \quad . \tag{12}$$

Inhomogeneous dielectrics for which Eqs. (9) and (12) can be satisfied will give rise to polarization independent propagation by non-coupled TE and TM fields.

When the spatial variation of the dielectric constant can be expressed in terms of a one-dimensional variable, Eqs. (9) and (12) can be reduced to explicit necessary and sufficient conditions for polarization independent propagation of non-coupled TE and TM modes.

Consider an orthogonal, curvilinear coordinate system $[\xi_1, \xi_2, \xi_3]$ with scale factors $[h_1, h_2, h_3]$ given by $h_j^2 = (\partial x/\partial \xi_j)^2 + (\partial y/\partial \xi_j)^2 + (\partial z/\partial \xi_j)^2$, $[x, y, z]$ being the Cartesian coordinates.[3] Let ϵ be a function of one coordinate only. Since the numbering

of the coordinate system is arbitrary, we choose this to be the ξ_1 coordinate, that is, $\varepsilon = \varepsilon(\xi_1)$. Moreover, we choose the direction of propagation to coincide with the direction of ξ_3, that is, we set $E_3' = 0$ and $H_3'' = 0$. It is then found from Eqs. (9) that for non-coupled TE and TM modes,

$$\underset{\sim}{E} = [0, E_2, 0] \quad , \quad \underset{\sim}{H}'' = [0, H_2, 0] \quad , \tag{13}$$

$$\frac{\partial}{\partial \xi_2}(h_1 h_3 E_2) = 0 \quad , \quad \frac{\partial}{\partial \xi_2}(h_1 h_3 h_2) = 0 \quad , \tag{14}$$

and Eqs. (1) and (12) reduces to the requirement that $\partial E_2'/\partial \xi_2 = 0$, $\partial H_2''/\partial \xi_2 = 0$,

$$\frac{\partial h_1}{\partial \xi_2} = 0 \quad , \quad \frac{\partial h_3}{\partial \xi_2} = 0 \quad , \quad \frac{\partial}{\partial \xi_2}\left(\frac{1}{h_2}\frac{\partial h_2}{\partial \xi_1}\right) = 0 \quad , \quad \frac{\partial}{\partial \xi_2}\left(\frac{1}{h_2}\frac{\partial h_2}{\partial \xi_3}\right) = 0 \quad , \tag{15}$$

with $\varepsilon(\xi_1)$ given by

$$\frac{d^2(1/\sqrt{\varepsilon})}{d^2\xi_1^2} + \left[\frac{h_1 h_2}{h_3}\frac{d}{d\xi_1}\left(\frac{h_3}{h_1 h_2}\right)\right]\frac{d(1/\sqrt{\varepsilon})}{d\xi_1} = 0 \quad . \tag{16}$$

Equation (15) specifies the coordinate systems and field polarizations for which polarization independent propagation of non-coupled TE and TM fields is possible; the dielectric constant profile $\varepsilon(\xi_1)$ which can support such fields is given by Equation (16).

For the plane-stratified case we choose a cartesian coordinate system (x, y, z) and assume an electromagnetic wave propagating in the z direction. With $[x, y, z] = [\xi_1, \xi_2, \xi_3]$ and $[h_1 = 1, h_2 = 1, h_3 = 1]$, we find from Eqs. (15) and (16) that the medium will support polarization independent propagation of non-coupled TE and TM modes, with $\underset{\sim}{E}_{TE} = [0, E_y, 0]$ and $\underset{\sim}{H}_{TM} = [0, H_y, 0]$, if there is no field variation in the y direction and

$$\varepsilon(x) = \varepsilon(0)\left(\frac{a}{x+a}\right)^2 \quad , \tag{17}$$

a being an arbitrary constant.

For rotationally symmetric cylindrical media, we consider a waveguide with a medium having a dielectric constant $\varepsilon(r)$ varying in the radial direction of a cylindrical coordinate system (r, ϕ, z). Assuming a wave propagating in the z direction, we have $[r, \phi, z] = [\xi_1, \xi_2, \xi_3]$ and $[h_1 = 1, h_2 = r, h_3 = 1]$, and Eqs. (15) and (16) show that the waveguide will support polarization independent non-coupled TE and TM modes, with $\underset{\sim}{E}_{TE} = [0, E_\phi, 0]$ and $\underset{\sim}{H}_{TM} = [0, H_\phi, 0]$, if the field is rotationally symmetric (i.e., $\partial/\partial\phi = 0$) and

$$\varepsilon(r) = \frac{\varepsilon(0)}{(1+\alpha^2 r^2)^2} \quad , \tag{18}$$

α being an arbitrary constant.

We observe that for practical optical communication fibers, the dielectric constant profile as given by Eq. (18) closely approximates the parabolic distribution $\varepsilon(r) = \varepsilon(0)(1 - a^2 r^2)$. Consequently, also in parabolic fibers, the propagation of rotationally symmetric TE and TM signals is nearly polarization independent. A comparison of the properties of the distribution as given by Eq. (18) with the parabolic distribution will be presented elsewhere.

We next consider a medium with a dielectric constant $\varepsilon = \varepsilon(\theta)$ varying in the θ direction of a spherical coordinate system (r, θ, ϕ). For a wave propagating in the radial direction, we have $[r, \theta, \phi] = [\xi_3, \xi_1, \xi_2]$ and $[h_3 = 1, h_1 = r, h_2 = r \sin\theta]$, and we will have polarization independent propagation of non-coupled TE and TM modes, with $\underset{\sim}{E}_{TE} = [0, 0, E_\phi]$ and $\underset{\sim}{H}_{TM} = [0, 0, H_\phi]$, if there is no field variation in the ϕ direction and

$$\varepsilon(\nu) = \frac{\varepsilon(0)}{(1 + A\cos\theta)^2} \quad , \tag{19}$$

A being an arbitrary constant.

The electromagnetic field of an infinity-strip laser resonator with parabolic end reflectors can be conveniently described in an elliptic cylinder coordinate system (u, v, z), where $1 \leq u < \infty$, $-1 \leq v \leq +1$, $-\infty < z < \infty$, surfaces u = constant are elliptic cylinders $x^2/u^2 + y^2/(u^2 - 1) = a^2$, a being a scale factor, surfaces v = constant are hyperbolic cylinders $x^2/v^2 - y^2/(1 - v^2) = a^2$, and surfaces z = constant are planes. According to Eq. (15), polarization independent propagation of non-coupled TE and TM modes if possible in such a coordinate system if $\underset{\sim}{E}_{TE} = [0, 0, E_z]$, $\underset{\sim}{H}_{TM} = [0, 0, H_z]$, and the dielectric constant ε varies with either the u or the v coordinate. Assuming that $\varepsilon = \varepsilon(u)$, we have $[u, v, z] = [\xi_1, \xi_3, \xi_2]$ and

$$\left[h_1 = a\sqrt{(u^2 - v^2)/(u^2 - 1)} \quad , \; h_3 = a\sqrt{(u^2 - v^2)/(1 - v^2)} \quad , \; h_2 = 1\right],$$

and the polarization independent profile is given by

$$\varepsilon(u) = \frac{\varepsilon(0)}{(1 + A\cosh^{-1}u)^2} = \frac{\varepsilon(0)}{[1 + A\,\ell n(u + \sqrt{u^2 - 1})]^2} \quad , \tag{20}$$

A being an arbitrary constant. For values of u close to unity Eq. (20) reduces to

$$\varepsilon(u) = \varepsilon(x, y) \approx \varepsilon(0)[(a^2 - x^2)/(\sqrt{a^2 - x^2} + A|y|^2] \quad .$$

Joint Services Technical Advisory Committee
F44620-75-C-0056

L. Bergstein

REFERENCES

1. J.A. Stratton, "Electromagnetic Theory," (New York: McGraw-Hill Book Co., 1941).

2. L. Bergstein and H. Oksman, "On Polarization-Independent Waveguides," Progress Report No. 38 to JSTAC, Polytech. Inst. of New York, Report No. R-452.38-73, pp. 162-167 (1973).

3. P.M. Morse and H. Feshleach, "Methods of Theoretical Physics, Part I," (New York: McGraw-Hill Book Co., 1953).

LEAKAGE AND RESONANCE EFFECTS ON STRIP WAVEGUIDES FOR INTEGRATED OPTICS

S. T. Peng and A. A. Oliner

A. Introduction

The basic strip waveguide for integrated optics which is under consideration is shown in Figure 1. As seen, this waveguide is not planar, but is a three-dimensional structure which confines the wave horizontally as well as vertically. The strip produces a central region with an effective refractive index greater than that of the surrounding regions; as a result, the surface wave that is present on the thin film is pulled in by the strip region and confined to its neighborhood. The guiding process is usually viewed in terms of waves in the strip region bouncing back and forth at an angle between the two sides of the strip, undergoing total reflection at each bounce.

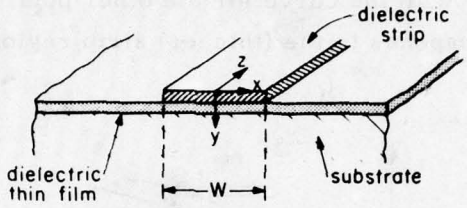


Fig. 1. The optical strip waveguide.

Theoretical analyses for the propagation characteristics of the strip waveguide have appeared in the literature,[1-4] but these analyses assume the presence of only one mode type, TE or TM, in each of the transverse regions comprising the cross section of the strip waveguide. When the field behavior at the strip sides is viewed more carefully, however, it is easy to see that a TE or TM mode incident on a strip side produces not only a reflected and transmitted wave of its own type, but also excites a reflected and a transmitted wave of the other type, plus a continuous spectrum (which is purely reactive here). This coupling at the strip sides between TE and TM modes produces the interesting propagation effects described here.

To show why such cross-coupling must occur, consider first a TE mode propagating along an infinitely-wide planar structure identical to that in the strip region. If the wave is propagating in the z direction, it has only the following field components: H_y, H_z and E_x. If, however, the wave travels at a small angle with respect to the z direction, additional small components H_x and E_z arise. Similarly, a TM mode traveling at a small angle in that configuration will possess not only components E_y, E_z and H_x, but also small amounts of E_x and H_z. When such a TE wave bounces back and forth in the actual strip region, it hits the strip sides at a small angle with respect to the z direction;

the small H_x and E_z field components then excite a TM mode at the strip sides. Similarly, the small E_x and H_z components of an incident TM mode will excite a TE mode there. Thus, TE-TM mode coupling necessarily occurs at these strip sides, and this small coupling gives rise to interesting consequences under suitable circumstances.

B. Qualitative Explanation of Leakage Effects

In order to obtain a simple qualitative understanding of the effects of TE-TM mode coupling on the guidance behavior, let us consider a special case of the strip waveguide, for which the strip material is the same as that of the thin film. Such a structure, referred to as a "rib waveguide" by Bell Laboratories, is shown in the inset in Figure 2. The dispersion curves presented in Fig. 2 apply to the constituent regions comprising the waveguide cross section, as if each were infinitely wide. The TE mode is seen to be the dominant mode, with the curve for the other polarization (TM) located nearby. The vertical line t_1 corresponds to the (thicker) strip region, and the t_2 line to the film region.

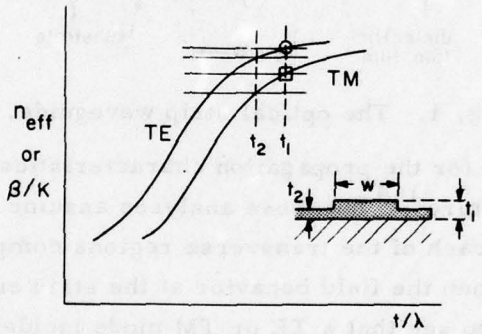


Fig. 2. Dispersion curves for the constituent waves in the strip and film regions.

For complete guiding, the wave or waves inside the strip region must be above cutoff (propagating) and the wave or waves outside must be below cutoff (evanescent). Suppose that the basic wave is TE, shown as a circle in Fig. 2 at thickness t_1, in the strip region. That point corresponds to the highest value of n_{eff}, and is therefore the slowest constituent wave. For guidance, the TE mode at thickness t_2, outside, must be evanescent. Thus, the TM modes which are excited in both the t_1 and t_2 regions are seen to be waves that are faster than the TE mode outside, and they are therefore certainly evanescent. Hence, for the lowest TE mode nothing interesting happens; the wave is still purely bound, although the numerical value of the propagation constant changes a bit due to the coupling.

When the basic mode is TM, however, we consider the square in Fig. 2 at thickness t_1. For guiding, that constituent wave must be propagating and the TM wave outside

(at t_2) must be evanescent. Thus, if the TM-TE mode coupling were neglected, the wave would be completely bound. The TE modes which are excited at the strip sides are, however, seen from Fig. 2 to be slower than the TM wave inside the strip, and thus to be above cutoff both inside the strip (t_1) and outside in the film region (t_2).

This situation is depicted in Figure 3. A TM mode incident from the inside excites a reflected propagating TM mode inside, an evanescent TM mode outside, and propagating TE modes both outside and inside. The propagating TE mode outside carries away a small amount of energy at each reflection resulting in a leaky wave rather than a purely bound wave.

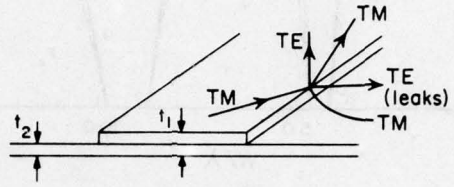


Fig. 3. TE-TM coupling in the "TM" case, showing leakage.

It is shown later that the propagating TE mode inside the strip region, since it is slower than the TM mode there, produces resonance effects.

Depending on how close the resonance curves are to each other, the TE mode outside may or may not be above cutoff. Thus, the TM mode may or may not be leaky, depending on the properties of the waveguide, but leakage will occur for many of the strip waveguides being considered today. Similar considerations apply to all higher modes, both TE and TM; leakage will occur under appropriate conditions.

C. Numerical Examples

Numerical examples are important in order to determine whether the leakage described above is negligible or significant. In the examples discussed below it is significant.

Let us consider the structure shown in the inset in Fig. 4, where the dimensions and refractive indices are indicated. Calculations were performed taking the TE-TM mode coupling into account correctly, assuming that only the dominant modes of each type can propagate in the constituent regions (valid for the film thicknesses chosen), and neglecting the continuous spectrum. It is found that the dispersion curves (n_{eff} vs. W/λ) are affected very little by the coupling. No curves are shown here, therefore.

The curve of attenuation vs. W/λ, on the other hand, is very interesting, and is shown in Figure 4. It should first be recognized that if the TE-TM coupling did not

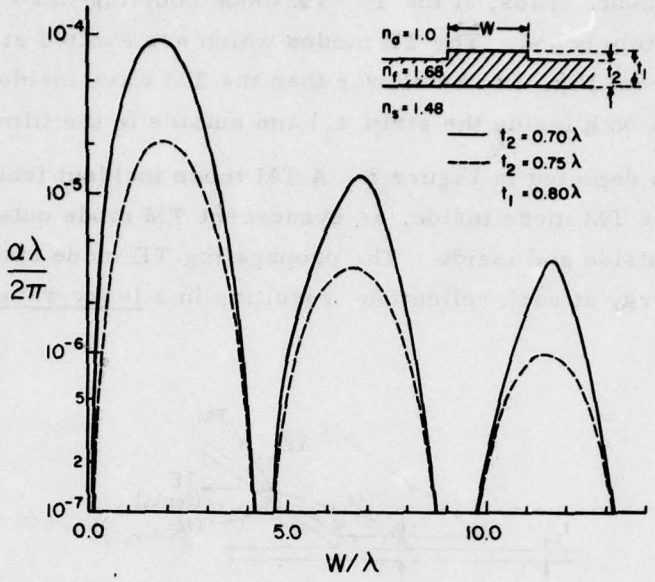


Fig. 4. Attenuation, due to TE-TM coupling, as a function of strip width for the strip waveguide.

occur there would be no attenuation, since we are neglecting material losses. Therefore, all the attenuation shown is due to leakage.

The numerical values for α are of the order of the measured results (including material losses), so that the leakage effect is significant. The leakage is seen to be greater for greater rib height, as expected since then a larger step discontinuity occurs at the strip side. The leakage also decreases as the strip width W increases. This occurs because, as W increases, the constituent waves inside the strip region approach closer to glancing incidence, and the coupled components decrease in amplitude.

In addition, it is evident from Fig. 4 that a resonance (or a cancellation) effect is occurring. It is due to the TE mode inside which is bouncing back and forth above cutoff. From a simple calculation, we find that to within 1%:

$k_x^{TE} W = 2\pi$ for the first resonance

$k_x^{TE} W = 4\pi$ for the second resonance

We have also derived an equivalent network for the step junction discontinuity corresponding to the strip side, and we can readily explain the resonance behavior using this network.

Similar considerations apply to other waveguiding structures, including the slot waveguide, consisting of an uncoated strip region between metal-coated regions.[5-9] We

have performed a similar analysis for a typical slot waveguide, and the numerical results show that the resonance effect dramatically reduces not only any leakage produced, but also the attenuation due to the presence of the metal.

The leakage and resonance properties have evident implications for device performance when TM polarization is used. For example, the properties of directional couplers, which consist of two waveguides placed parallel to each other over a specified length, could be influenced in important ways. Because of the leakage, the coupled waveguides need not be as near to each other for the same coupling length. Additional flexibility is available since the leakage can be avoided altogether by making the strip side sufficiently high (see Figure 2). Then, the film thickness outside of the coupled region can be made small so as to avoid leakage outward, but the film thickness in the region between the two strips would be made greater to permit leakage to occur and thus to enhance the coupling between the strips. We are in the process of examining the effects of leakage on such directional coupler performance.

Joint Services Technical Advisory Committee
F44620-74-C-0056

S. T. Peng and A. A. Oliner

REFERENCES

1. H. Furuta, H. Noda and A. Ihaya, "Novel Optical Waveguide for Integrated Optics," Appl. Opt., Vol. 13, p. 322 (February 1974).
2. V. Ramaswamy, "Strip-loaded Film Waveguides," Bell Syst. Tech. J., Vol. 53, p. 697 (April 1974).
3. E.A.J. Marcatili, "Slab-Coupled Waveguides," Bell Syst. Tech. J., Vol. 53, p. 645 (April 1974).
4. N. Uchida, "Optical Waveguide Loaded with High Refractive-Index Strip Film," Appl. Opt., Vol. 15, p. 179 (January 1976).
5. A.A. Oliner, "Optical and Acoustical Micro-Electronics: Similarities and Differences," in Proc. Symp. Optical and Acoustical Micro-Electronics (New York: Polytechnic Press, 1974) Vol. XXIII, p. 1.
6. J. Hamasaki and K. Nosu, "A Partially Metal-Clad-Dielectric-Slab Waveguide for Integrated Optics," IEEE Trans. Quantum Electron., Vol. 10, p. 822 (October 1974).
7. Y. Yamamoto, T. Kamiya and H. Yanai, "Propagation Characteristics of a Partially Metal-Clad Optical Guide: Metal-Clad Optical Strip Line," Appl. Opt., Vol. 14, p. 322 (February 1975).
8. A.A. Oliner and S.T. Peng, "Optical Slot and Strip Waveguides Formed by Metal Platings," Digest of Topical Meeting on Integrated Optics, Paper No. MC5, Salt Lake City, Utah (January 1976).
9. A.A. Oliner and S.T. Peng, "Waveguides for Integrated Optics Formed by Metal Platings," Proc. Sixth European Microwave Conf., p. 499, Rome, Italy (September 1976).

A NOVEL LEAKY-WAVE STRIP WAVEGUIDE DIRECTIONAL COUPLER FOR INTEGRATED OPTICS

E-W. Hu, S.T. Peng and A.A. Oliner

A. Introduction

A new type of directional coupler for integrated optics is presented here which looks like the ordinary strip waveguide type of directional coupler but which operates on a different coupling mechanism. As shown in Fig. 1, the coupler consists of two strip waveguides placed on a thin film on a substrate, where the two strips couple to each other over a given length. However, its operation allows a greater separation between the two strips than is present in the customary coupler.

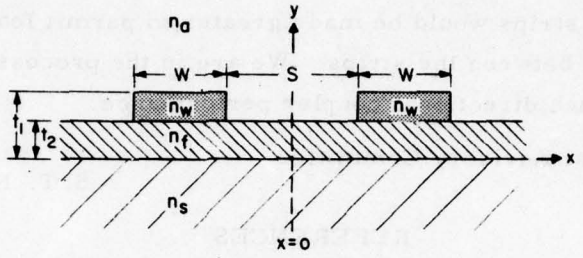


Fig. 1. Cross section of the directional coupler consisting of two strips, spaced s apart, on a thin film on a substrate.

In the ordinary type of strip waveguide coupler, each strip lies in the exponentially-decaying transverse field of the other strip, and the coupling which occurs is due to these fields. The coupling is a very sensitive function of the separation (s in Fig. 1) between the strips, and the coupling length, say the distance along the strips for complete transfer of power, is therefore sensitively dependent on this separation. If the separation is too great, of course, the coupling becomes negligible and the coupling length becomes excessive.

The new type of directional coupler proposed here depends on a new phenomenon, leakage between the strips which arises under appropriate design conditions. Because of this leakage, the coupling is no longer as sensitive to the separation s, and the coupling length can be adjusted to the desired length, within limits, by controlling the amount of leakage. Since the separation s need no longer be very small, and since the coupling length is less sensitively dependent on it, this new type of coupler relaxes the fabrication tolerances in certain ways.

As will be shown below, leakage occurs only for the mode for which the electric field is vertical. That mode corresponds to basic TM excitation (all the modes are

actually hybrid, of course). The basic TE mode does not leak. The directional coupler is therefore restricted in its operation to the TM-type mode. On the other hand, this restriction can be a virtue since it permits coupling for the TM-type mode but not for the TE-type mode if the separation s is sufficiently large (a half-dozen wavelengths or so, depending on the parameters).

In the discussions which follow, the origin of the leakage is explained, the operation of the coupler is described, and numerical results obtained from a theoretical analysis are presented for a specific example.

B. Explanation of the Leakage Effect

Let us consider first the propagation behavior of a single isolated strip on the film and substrate. The guiding process is usually viewed in terms of waves in the strip region bouncing back and forth at an angle between the two sides of the strip, undergoing total reflection at each bounce. The approximate theoretical analyses which have appeared in the literature[1-4] assume the presence of only one mode type, TE or TM, in each of the transverse regions comprising the cross section of the strip waveguide. When the field behavior at the strip sides is viewed more carefully, however, it is easy to see that a TE or TM mode incident on a strip side produces not only a reflected and transmitted wave of its own type, but also excites a reflected and transmitted wave of the other type, plus a continuous spectrum (which is purely reactive here). It is this small coupling at the strip sides between TE and TM waves that produces the leakage which forms the basis for the novel directional coupler.

This leakage effect was recognized and described recently by two of the present authors.[5] All of the previously-published analyses miss this effect entirely.

A discussion of the leakage effect, and a qualitative explanation regarding when it occurs and how it manifests itself, are included in the report[6] immediately preceding this one. That explanation made use of a special case of the strip waveguide, for which the strip material is the same as that of the thin film. Such a structure has been called the "rib waveguide" by the Bell Laboratories in their studies.

It is shown in the preceding report[6] that a TM mode incident on a strip side from the inside of the strip excites a reflected propagating TM mode inside, an evanescent TM mode outside, and propagating TE modes both outside and inside. The propagating TE mode outside carries away a small amount of energy at each reflection, resulting in a leaky wave rather than a purely bound wave. Depending on how close the resonance curves of the constituent regions are to each other, the TE mode outside may or may not be above cutoff. Thus, the TM mode may or may not be leaky, depending on the properties of the waveguide.

C. The Operation of the Leaky-Wave Coupler

We may now return to the directional coupler configuration of Figure 1. When one of the two strip waveguides is excited in the basic TE mode (vertical magnetic field), that waveguide does not leak, and the other strip waveguide senses only its transversely evanescent field. The directional coupler configuration then operates in standard fashion, so that negligible coupling between the strips would occur if the strip separation s were made sufficiently large.

On the other hand, when the strip is excited in the basic TM mode (vertical electric field), the wave may leak if the geometrical parameters and refractive indices are selected appropriately. The field components present in the film region under these conditions of leakage consist of the transversely evanescent TM wave and the propagating TE wave, the latter possessing small amplitude. Actually, the propagating TE wave also slowly increases exponentially in the transverse (x) direction, consistent with the small exponential decay in the longitudinal (z) direction of the leaky guided wave, but this property affects the performance negligibly. When the strip separation s is large, the other strip no longer senses the evanescent TM field but it couples to the transversely propagating TE wave component.

The performance of the coupler was analyzed (taking both the TE and TM components into account) by determining the propagation characteristics of the symmetric and the antisymmetric modes of the overall structure comprised of both strips. The symmetric and antisymmetric modes behave slightly differently even when the separation s is large because the leaky TE wave component reacts differently to the electric or magnetic wall nature of the midplane at $x = 0$.

Numerical values for a typical special case are shown in Fig. 2 for both basic TE and basic TM excitation. Plotted are $n_{eff}(=\beta/k=\lambda/\lambda_g)$ as a function of s/λ for both the symmetric and antisymmetric modes. It is seen that for the TE basic wave, which does not leak, the behavior is the customary one. For the leaky TM basic wave, one notes that the coupling continues even as the separation s becomes large. In fact, the symmetric and antisymmetric mode contributions interlace with each other, and the regions of maximum spacing between them (which occur periodically) are relatively flat.

The coupling length L, which can be defined as the length corresponding to complete power transfer, is given by $\pi/|\beta_s - \beta_a|$, where β_s and β_a are the propagation wavenumbers of the symmetric and antisymmetric modes, respectively. For a strip separation $s = 12\lambda$, one finds $L \approx 2500\lambda$, which is a desirable length. Furthermore, one sees from Fig. 2 that a small fabrication error in the value of s will negligibly affect

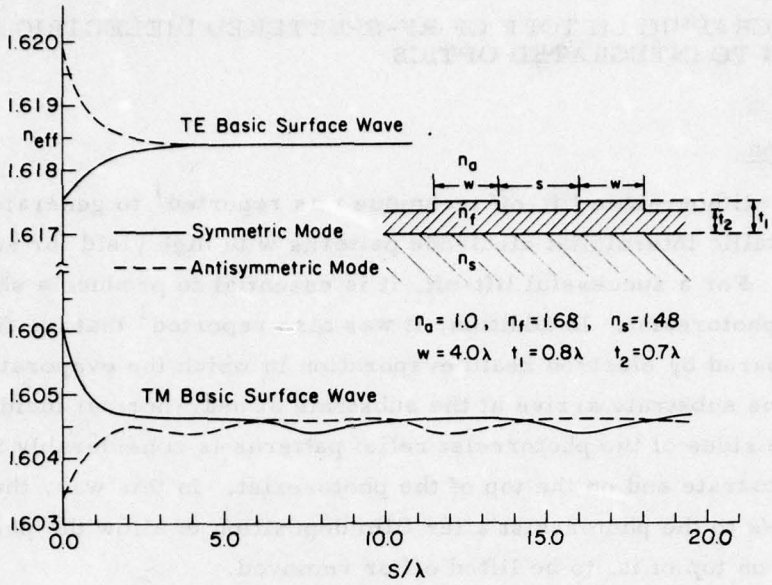


Fig. 2. Coupling properties of the leaky-wave directional coupler. Shown are $n_{eff}(=\beta/k=\lambda/\lambda_g)$ as a function of s/λ, where s is the strip separation, for both symmetric and antisymmetric modes on the coupler structure, for both basic TE and TM excitations.

the value of L, so that this type of coupler should permit looser tolerances in some parameters.

In the structure of Fig. 1, leakage also occurs in the film regions outside the strips. If such leakage is undesirable, it can be eliminated by making the film thicker in the region between the strips than in the outside regions. Then one can achieve the desired leakage between the strips and yet avoid leakage outward. The basis for this statement follows from the discussion in the preceding report.[6]

Joint Services Technical Advisory Committee
F44620-74-C-0056

A.A. Oliner

REFERENCES

1. H. Furuta, H. Noda and A. Ihaya, "Novel Optical Waveguide for Integrated Optics," Appl. Opt., Vol. 13, p. 322 (February 1974).
2. V. Ramaswamy, "Strip-Loaded Film Waveguides," Bell Systems Tech. J., Vol. 53, p. 697 (April 1974).
3. E.A.J. Marcatili, "Slab-Coupled Waveguides," Bell Systems Tech. J., Vol. 53, p. 654 (April 1974).
4. N. Uchida, "Optical Waveguide Loaded with High Refractive-Index Strip Film," Appl. Opt., Vol. 15, p. 179 (January 1976).
5. S.T. Peng and A.A. Oliner, Tech. Digest of the Intl. Conf. on Integrated Optics and Optical Fiber Communication (IOOC '77), pp. 29-32, Tokyo, Japan (July 18-20, 1977).
6. S.T. Peng and A.A. Oliner, "Leakage and Resonance Effects on Strip Waveguides for Integrated Optics," Progress Report No. 40 to JSTAC, Polytech. Inst. of New York, Report No. R-452.42-77 (1977).

PHOTOLITHOGRAPHIC LIFTOFF OF RF-SPUTTERED DIELECTRIC FILMS, WITH APPLICATION TO INTEGRATED OPTICS

E-W. Hu

A. Introduction

The photolithographic lift-off technique was reported[1] to generate one micrometer line-width metallic interdigital electrode patterns with high yield for surface acoustic wave devices. For a successful lift-off, it is essential to produce a sharp vertical profile in the photoresist. In addition, it was also reported[1] that the film should preferably be prepared by electron beam evaporation in which the evaporated atoms to be deposited on the substrate arrive at the substrate at near-normal incidence such that the film on the sides of the photoresist relief patterns is considerably thinner than the film on the substrate and on the top of the photoresist. In this way, the organic solvent can have access to the photoresist after film deposition to allow the photoresist, along with the films on top of it, to be lifted off or removed.

Besides the application of the lift-off technique to metallic films, Wolf[2] demonstrated its applicability to ion beam sputter-deposited glass films on patterns generated by a scanning electron microscope for the construction of optical waveguides. However, for some unknown reasons,[2] the edges of the waveguides thus fabricated were not smooth and sharp. As to rf-sputtered films, it is often believed that the sputtered atoms or ions possess poor directionality and hence make the lift-off difficult. It was briefly reported that, even in the case of very thin metal films (300-500 Å of Chromium), the unwanted metal films could not be thoroughly removed.[1] However, it seems that no further detailed work along this line has been reported in the literature.

Rf-sputtering has been extensively used for making thin films for various applications. It seems particularly attractive for the preparation of low-loss films for integrated optics applications. Despite the complex plasma processes associated with sputtering, the important sputtering parameters can be accurately controlled to yield reproducible results. Furthermore, deposition parameters such as gas content and sputtering powers can be readily adjusted over a wide range. This, combined with a relatively slow deposition rate, offers the possibility of depositing films with accurate control of thickness and of variable refractive index.

In this work, we report our successful application of the photolithographic lift-off technique to rf-sputtered thin dielectric films and to the fabrication of optical strip-line waveguiding structures.[3]

The optical strip-line waveguide is one of the most important components in integrated optics. It consists, as shown in Fig. 1, of a basic thin film waveguide loaded

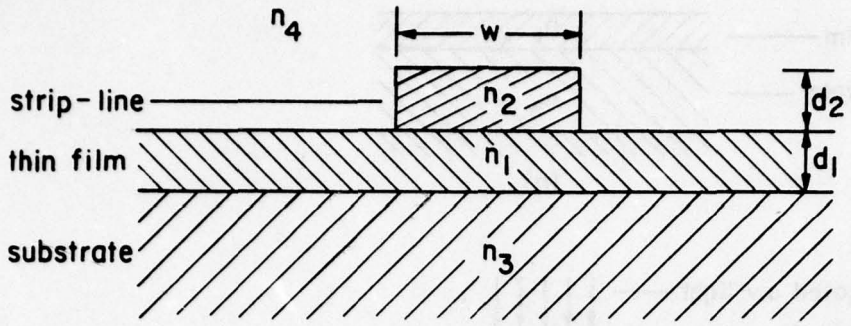


Fig. 1. Optical strip-line waveguiding structure: $n_1 \geq n_2 > n_3 > n_4$ or $n_2 \geq n_1 > n_3 > n_4$, where n_1, n_2, n_3 and n_4 are the refractive indices of film, strip-line, substrate and air, respectively.

by a dielectric strip of finite width W. In the strip region, the guided surface wave is slower than that in the other regions. As a result, the surface wave energy becomes trapped in the vicinity of the strip and can therefore be guided one-dimensionally. This may provide a method for direct coupling to an optical fiber because the wave is confined transversely in both directions and offers other advantages over the two-dimensional (planar) simple waveguide. Also, by setting $n_1 > n_2$ (Fig. 1), most of the energy carried by the surface wave is confined in the thin film beneath the strip; therefore, the requirement for very smooth edges for low-loss waveguiding is less stringent than that for other types of linear optical waveguide. Moreover, the optical strip line is the building block for more sophisticated integrated optical devices, such as modulators,[4] directional couplers and switches,[5] etc. Currently, there are several techniques for making optical strip line and similar structures.[6] However, except for the one that was mentioned before,[2] most of these techniques make use of either ion beam or rf-etchings; the procedures involved were often tedious and complex. As will be discussed later, the significance of our current approach is that it readily provides an alternate and a simple way for the fabrication of optical strip-line structures. The experimental procedures are discussed in detail in the following section.

B. Experiments

The main steps of the fabrication process are illustrated in Fig. 2 and are discussed below. As shown in Fig. 2(a), the basic surface waveguide is first made by rf-sputtering 7059 glass or an Al_2O_3 film on a pre-cleaned 1" x 3" microscope slide. The thickness and refractive index of the film were measured by the prism-coupler method. The facilities and procedures for film preparation and characterization were discussed in our previous work.[7] Both the refractive index and the thickness of the film

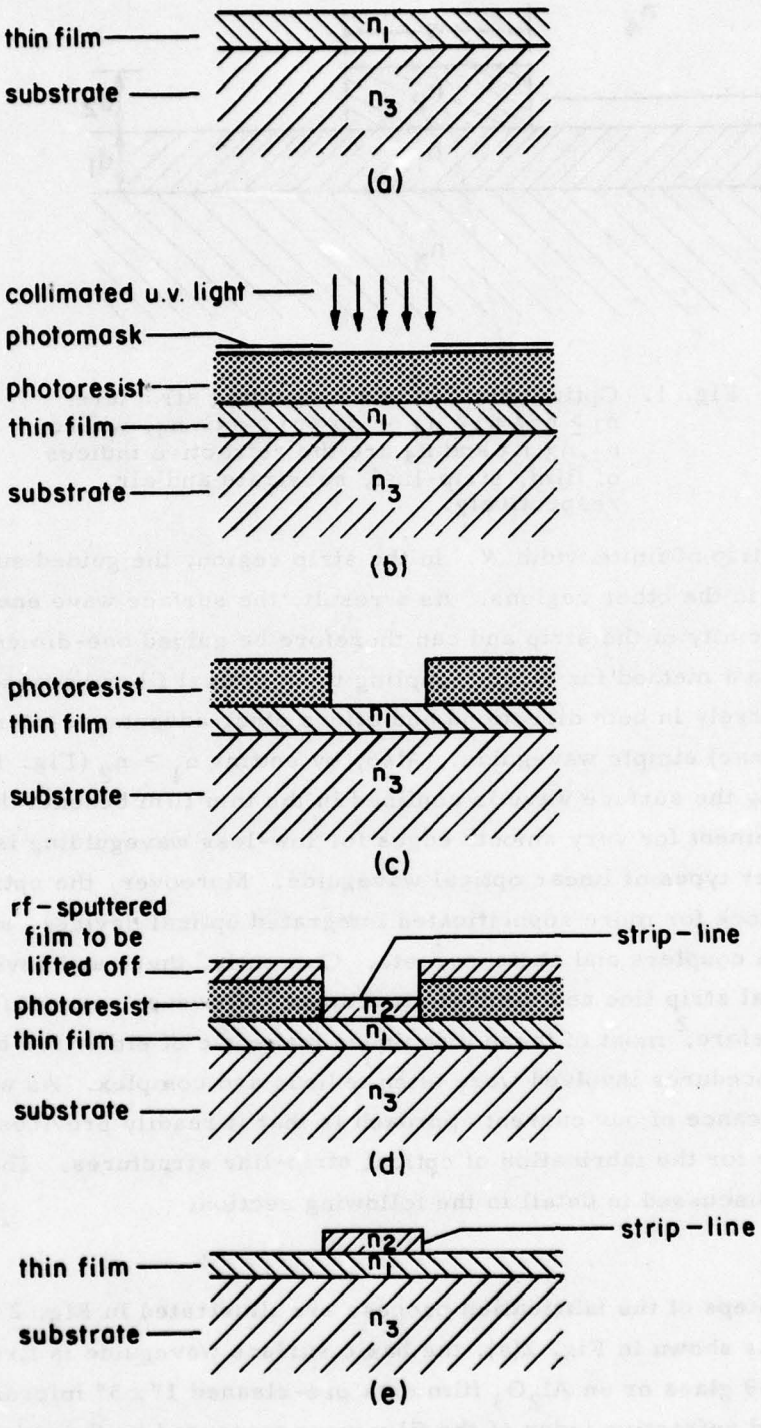


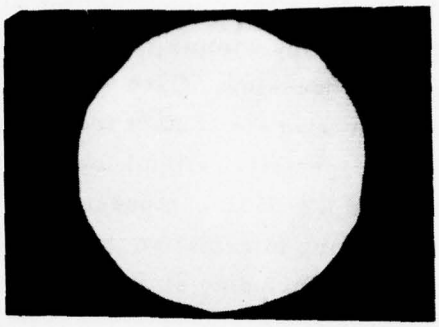
Fig. 2. Main steps of photolithographic lift-off technique for the fabrication of optical strip-line waveguide.

were accurately controlled such that the film guided only the dominant TE and TM modes. Following the film deposition, the sample was immediately transferred to the laminar hood for photolithographic processing. Care was exercised to avoid airborne dusts which would, later on, prevent intimate contact from forming between the photomask and the photoresist. This is especially critical for patterns with micrometer or submicrometer dimensions. Shipley AZ1350B photoresist was chosen in this work for the availability of most of its processing information through previous work done here.[8] Standard photolithographic procedures including spin-coating, baking, exposing and subsequent developing of the photoresist were performed (Figures 2(b) and 2(c)). A collimated ultra-violet exposure beam of uniform intensity was derived from a 200 W high pressure Hg light bulb. Intimate contact between the photomask and the sample was checked by observing interference fringes.[1] With intimate contact between mask and sample, the exposure and development time were found to be not critical.[1] Therefore, the sample was always over-exposed and over-developed to insure the complete removal of the exposed photoresists.

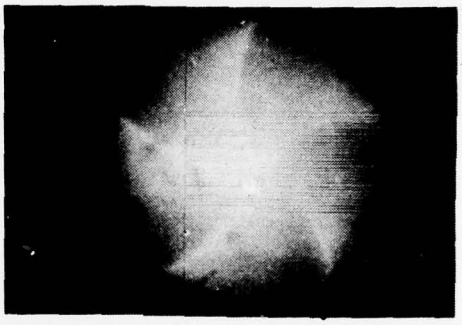
After development of the photoresists, the sample was baked at 50°C for 20 minutes in a vented oven. The sample was then transferred to the vacuum chamber for strip-line film deposition. Since the softening point of the photoresist is rather low (below 150°C), no pre-etch of the substrate was done. The sputter-deposition of the strip (Fig. 2(d)) was performed with rf-sputtering powers below 100 W for most cases. Although the base-plate of the rf-system employed in this work was water cooled, it was found that for sputtering powers higher than 100 W, deformation and even blistering of the photoresists would occur.

Following the strip deposition, the unwanted photoresists and films were lifted off by dipping the sample in acetone for about 20 minutes (Fig. 2(e)). Gentle swabbing of the sample by an acetone-saturated cotton Q-tip was occasionally needed to remove the residual photoresist, depending on the resist-to-film thickness ratio. For a 4:1 ratio, no difficulty was encountered in the lift-off procedure. Figure 3(a) shows the photomicrograph of the photoresist relief pattern of an array of optical strip lines with resist thickness of about 10,000 Å. Figure 3(b) shows the actual glass strip-line structure after lift-off. The strip height in Fig. 3(b) is 3,000 Å and the strip width 7.5μ. The separation distance between strips is 5μ. Comparison between the photoresist relief pattern and the actual waveguide structure (Figs. 3(a) and 3(b)) shows that they are almost identical, implying the accurate mapping of the photomask patterns onto the resist and the subsequent strip-line structure.

However, the actual smoothness of the sides of the strips must be checked with a scanning electron microscope. Such pictures have been gratifying in that they show



(a)



(b)

Fig. 3. Photomicrographs of
(a) Photoresist relief pattern of an array of strip-lines, where the resist thickness is 10,000 Å.
(b) Glass optical strip-line waveguide obtained from the above photoresist relief pattern by the lift-off technique, where the strip width = 7.5 μ, the separation distance between strips = 5 μ and the strip height = 3,000 Å.

very sharp and steep strip sides. Some inadequacies still appear, however, and they will require further investigation. In some places, ripples appear on the uncoated glass substrate, indicating that the substrate should be optically polished before use. Sometimes, the top edge of the strip side is rounded somewhat; it is believed that this is due to the swabbing by the Q-tip, but this point will be checked further. Finally, although the strip sides are satisfactorily vertical, they ripple a bit along their lengths. It is known that this ripple is due to the mask used, which was kindly furnished to us by the MIT Lincoln Laboratory.* Although certain improvements are called for, there-

*A chromium-coated thin glass (Corning 0211) comformable photomask with interdigital transducer patterns originally designed for surface acoustic wave device applications.

fore, the results so far are very encouraging, and they do indicate the success of the method in producing the required verticality in the strip sides.

C. Conclusions

We have found that the lift-off technique can be used successfully with rf-sputtered dielectric films. Our method therefore provides a convenient way to fabricate optical strip lines and similar structures for integrated optics applications. Even though we used a photomask pattern with a line width of a few micrometers, the results imply that the dimensions could be reduced to a micrometer or less. This is because the resolution of the line width is limited by the photolithographic technique rather than by the lift-off technique. As has been reported in a review article by Smith,[9] the comformable photomasking technique could yield well-defined lines as narrow as a few thousand angstroms in photoresists with near vertical profile and very little undercut.

As discussed in the previous section, the ratio of the thickness of the photoresist to that of the film should be approximately 4: 1 for easy lift-off. For Shipley AZ1350B resist used in this work, it is difficult to obtain uniform resist layers thicker than about 1.6 μm. This limits the film to be lifted off to about 4,000 Å. For thicker films or thicker strips, other types of photoresist must be used. Fortunately, there are a number of commercially-available high-resolution photoresists which could be deposited as thick as a few micrometers.

Besides the optical strip-line structure discussed in this work, the fabrication procedures should also be valuable in making optical grating structures. So far, it seems that most of the gratings used as input/output beam couplers for integrated optics applications were made by recording the laser-beam interference patterns in photoresists which are quite lossy at optical frequencies.[8] Because of the lossy nature of the photoresists, the difficulty in precisely controlling the grating profile, and other problems associated with holographical grating generation, it seems that there is still a poor correlation between theory and actual device performance of the grating couplers. Since almost all the important fabrication parameters of our current approach could be accurately controlled, and the periodic grating structure could be made of low-loss dielectric film, our experimental procedures may provide a simple means for a systematic experimental study of the grating couplers.

Currently, work is underway using this approach to construct a specially-designed directional coupler[11] to verify a newly-discovered leakage phenomenon associated with the basic strip-line waveguide,[10] and to determine whether or not this leaky-wave directional coupler is indeed much less sensitive to dimensional tolerances.[11]

Joint Services Technical Advisory Committee
F44620-74-C-0056

E-W. Hu

REFERENCES

1. H.I. Smith, F.J. Bachner and N. Efremow, "A High-Yield Photolithographic Technique for Surface Wave Devices," J. Electrochem. Soc., Vol. 118, pp. 821-825 (May 1971).

2. M.K. Barnoski, "Introduction to Integrated Optics," (New York: Plenum Press, 1973), Ch. 7.

3. H. Furuta, H. Noda and A. Ihaya, "Novel Optical Waveguide for Integrated Optics," Appl. Optics, Vol. 13, No. 2, pp. 322-326 (February 1974).

4. I.P. Kaminow, V. Ramaswamy, R.V. Schmidt and E.H. Turner, "Lithium Niobate Ridge Waveguide Modulator," Appl. Phys. Lett., Vol. 24, No. 12, pp. 622-624 (June 1974).

5. S. Somekh, E. Garmire, A. Yariv and R.G. Hunsperger, "Channel Optical Waveguides and Directional Couplers in GaAs-Imbedded and Ridged," Appl. Optics, Vol. 13, No. 2, pp. 327-330 (February 1974).

6. M.K. Barnosky, "Introduction to Integrated Optics," (New York: Plenum Press, 1973), Ch. 6 and 7.

7. L. Bergstein, E-W. Hu and M. Eschwei, "Rf-sputtered Al_2O_3 Films for Integrated Optics Applications," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76 (1976).

8. S. Austin, "Periodic Integrated Optical Structures in Photoresist: Application to Grating Couplers," Ph.D. Dissertation, Polytech. Inst. of New York (1976).

9. H.I. Smith, "Fabrication Techniques for Surface-Acoustic-Wave and Thin-Film Optical Devices," Proc. IEEE, Vol. 62, No. 10, pp. 1361-1387 (October 1974).

10. S.T. Peng and A.A. Oliner, "Leakage and Resonance Effects on Strip Waveguides for Integrated Optics," Technical Digest, Conf. on Integrated Optics and Optical Fiber Communications (IOOC), Tokyo, Japan (July 1977).

11. E-W. Hu, S.T. Peng and A.A. Oliner, "A Novel Leaky-Wave Strip Waveguide Directional Coupler for Integrated Optics," Progress Report No. 42 to JSTAC, Polytech. Inst. of New York, Report No. R-452.42-77 (1977).

SUPERRADIANT OUTPUT FROM A DISCHARGE-HEATED COPPER VAPOR LASER

W. T. Walter

The copper vapor laser is one of the few high-power sources available in the blue-green spectral region, near the maximum of the water transmission band. Operation of the copper vapor laser differs from most other lasers in that a high temperature region is required to generate sufficient vapor density. Although the high temperature presents some difficult materials problems, it also offers the important advantage of allowing radiation to be used to stabilize the operating temperature. Radiation can remove the limitations imposed by thermal conduction and eliminate the required external water-cooling systems of present high-power systems. Demonstration of a practical discharge-heated system is important therefore, for the further development of this class of efficient, pulsed gas-discharge lasers.

The first copper vapor lasers[1-5] utilized furnaces to produce operating temperatures of 1400-1600°C. At Polytechnic 100-kW peak-power, 2-mJ pulses have been demonstrated with an 0.7%- energy-conversion efficiency. More than 10 W of average-power laser output has been generated using an externally-heated laser tube which required 2700 W of heater power.[5] Although the operating temperature can be reduced by using a compound source of copper, such as a copper halide,[6-9] the system efficiency is reduced because electrical discharge energy must be provided to dissociate several molecules for each copper atom excited through the laser cycle. In addition the operating regimes of copper halide systems are very narrow, particularly in terms of tube temperature or density of the copper halide vapor. Our approach has been to develop an optimized discharge-heated atomic-copper system by increasing the current density, discharge impedence and pulse repetition frequency and controlling the thermal conductivity so that the energy dissipated in the discharge will provide the optimum tube temperature. Recycling electrodes which utilize capillary action to confine the metal vapor are also being examined. Under these conditions the metal vapor laser is self-heated. No external furnace is required.

Previously a discharge-heated copper vapor laser[5] had been demonstrated at Polytechnic. It operated for short periods of time however, with an average power output of only 0.1 W. An improved discharge-heated copper vapor laser has now been constructed which can be operated for many hours. Tungsten mesh electrodes were spaced 55 cm apart inside a 12 mm i.d. boron nitride 3-section tube containing 18 gms of copper shot. The initial measurements reveal significant differences with the laser output from an externally-heated copper vapor laser. The operating mode of the discharge-heated device appears to be much closer to that of a superradiant, superfluorescent or amplified spontaneous emission generator. The presence or absence

of mirrors forming an optical resonator has a much smaller effect on the output, and the average output power is much lower than that from an externally-heated copper vapor laser. A 3-nsec pulse is present in the output of the discharge-heated laser without any mirrors as indicated in the upper trace of Figure 1. The addition of mirrors does not change the size of this pulse but appears to lengthen it by the addition of several subsequent pulses of similar amplitude. The middle and upper traces of Fig. 1 show the effect of adding first the far mirror and then both mirrors. Thus addition of mirrors changes the output power by a relatively small factor (2 to 4). The average output power has been less than ~0.5 watt thus far. The externally-heated devices, on the other hand, have very small superradiant outputs. When mirrors are added the output power increases by several factors of ten, and average power outputs as high as 11 W have been obtained.[5]

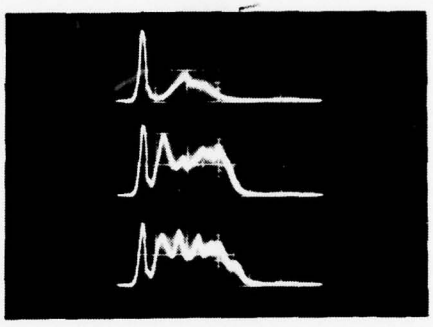


Fig. 1. Output at 5105 Å of discharge-heated copper vapor laser with 10 torr argon operating at 5680 pps. Detection system consisting of TRG 105B vacuum photodiode detector and Tektronix 519 oscilloscope has 1 GHz bandwidth and risetime < 0.5 nsec. Horizontal scale is 10 nsec/large division.
a) top trace - no optical resonator (some or all of the second peak appears to be produced by diffuse reflection back into the copper gain tube).
b) middle trace - Aℓ plane mirror installed at far end.
c) bottom trace - optical resonator completed 20%-partially-reflecting plane output mirror.

Since the discharge-heated configuration has a higher total system efficiency and is a more practical device form, an examination is being undertaken of the differences in operation and output. The aim is to obtain an understanding of the various processes involved as well as to find operating regimes which do not suffer from the deficiencies

revealed in these initial measurements. To compare the externally-heated and discharge-heated lasers, we must improve our temperature measuring procedures. A two-color, ratio-reading optical pyrometer will be examined for this purpose. The hottest region of the externally-heated laser is the o.d. of the tube on which the heater is wound, while the hottest region of the discharge-heated laser is the center of the discharge tube near the cathode. In addition in the discharge-heated device the composition and pressure of additive gases is more critical in keeping the impedance of the discharge substantially greater than the impedance of the thyratron switch.

Finally the question of superradiance or cooperative spontaneous emission behavior is also being examined. Many high-power lasers also possess high gain ($\alpha L > 1$) and therefore may be subject to limitations imposed by superradiant, superfluorescent or cooperative spontaneous emission effects. In these processes atoms are coupled together by their common radiation field and so radiate cooperatively. The intensity emitted in the phased direction by N electric dipole moments is therefore proportional to N^2 instead of N. An output intensity $\sim N^2$ requires that $\alpha L > 1$ where α is the intensity gain coefficient. According to Bonifacio and Lugiato[10] however, superfluorescence also requires that the length of the active medium must be longer than a threshold length α^{-1} characteristic of stimulated processes and shorter than a cooperation length characteristic of the exchange of energy between the field and the excited atoms. This latter requirement is disputed by MacGillivrey and Feld.[11] Experiments by Gibbs[12] in cesium vapor and experiments by others in HF. CH_3F and Na and Tl vapors have not resolved this question. Pulses generated by the copper vapor laser without mirrors are very similar in character and shape to those generated in cesium and other systems prepared to study cooperative emission effects. Cooperative emission behavior can be examined using the copper vapor laser which has the twin advantages of being able to generate larger inversion densities (up to $10^{14} cm^{-3}$) and having its output in the visible spectral region where detection systems are more sensitive and have wider bandwidths. A wider dynamic range is therefore available to examine these effects and determine their limitations on high-power efficient lasers.

Joint Services Technical Advisory Committee
F44620-74-C-0056

W. T. Walter

REFERENCES

1. W. T. Walter, N. Solimene, M. Piltch and G. Gould, "Efficient Pulsed Gas Discharge Lasers," IEEE J. Quantum Electron, Vol. QE-2, pp. 474-479 (1966).

2. W. T. Walter, "40-kW Pulsed Copper Vapor Laser," Bull. Am. Phys. Soc., Vol. 12, p. 90 (1967) and W. T. Walter, "High Power Copper Vapor Laser," Final Report NADC Contract N62269-3584 (1966).

3. W. T. Walter, "Improved Copper Vapor Laser," Progress Report No. 36 to JSTAC, Polytech. Inst. of Brooklyn, Report No. R-452.36-71, pp. 112-113 (1971).

4. W. T. Walter, N. Solimene, J. T. LaTourrette and K. Park, "Chemically-Generated Copper Vapor Lasers," Progress Report No. 38 to JSTAC, Polytech. Inst. of New York, Report No. R-452.38-73, pp. 121-128 (1973).

5. W. T. Walter, N. Solimene and J. T. LaTourrette, "Discharge-Heated Copper Vapor Lasers," Progress Report No. 40 to JSTAC, Polytech. Inst. of New York, Report No. R-452.40-75, pp. 177-178 (1975).

6. C. S. Liu, E. W. Sucov and L. A. Weaver, "Copper Superradiant Emission From Pulsed Discharges in Copper Iodide Vapor," Appl. Phys. Lett., Vol 23, pp. 92-93 (1973).

7. C. J. Chen, N. M. Nerheim and G. R. Russell, "Double-Discharge Copper Vapor Laser with Copper Chloride as a Lasant," Appl. Phys. Lett., Vol 23, pp. 514-515 (1973).

8. C. J. Chen and G. R. Russell, " High-Efficiency Multiply Pulsed Copper Vapor Laser Utilizing Copper Chloride as a Lasant," Appl. Phys. Lett., Vol. 26, pp. 504-505 (1975).

9. C. S. Liu, D. W. Feldman, J. L. Pack and L. A. Weaver, "Kinetic Processes in Continuously Pulsed Copper Halide Lasers," IEEE J. Quantum Electron, Vol. QE-13, pp. 744-751 (1977).

10. R. Bonifacio and L. A. Lugiato, "Cooperative Radiation Processes in Two-level Systems: Superfluorescence," Phys. Rev. A, Vol 11, pp. 1507-1521 (1975).

11. J. C. MacGillivray and M. S. Feld, "Theory of Superradiance in an Extended, Optically Thick Medium," Phys. Rev. A, Vol. 14, pp. 1169-1189 (1976).

12. H. M. Gibbs, Q. H. F. Vrehen and H. M. J. Hikspoors, "Superfluorescence in Cesium: Comparison with Theory and Application to Quantum Beat Spectroscopy," in Laser Spectroscopy, J. L. Hall and J. L. Carlsten, eds., (Springer-Series in Optical Sciences, Vol. 7), Springer-Verlag, N. Y., Heidelberg (1977).

COMPUTER MODELING AND EXAMINATION OF EXCITATION RATES FOR A BISMUTH VAPOR LASER AT 4722 Å

W. T. Walter and N. Solimene

Laser action at 4722Å in bismuth vapor was predicted to be a member of a class of efficient, pulsed gas-discharge lasers.[1] Laser action at this wavelength would be very close to the center of the water transmission band. Although a number of laboratories have examined discharges in bismuth vapor, laser action has not yet been reported.‡ Bismuth vapor has a substantial fraction (~ 50%) of dimers, Bi_2 molecules, which could prevent laser action by means of several absorption processes. Two methods of dissociating the dimers, thermal[2] and double-pulse discharge,[3] have been experimentally explored without achieving laser action. The spontaneous emission lifetime of the 4722Å $6p^2 7s\ {}^4P_{1/2} \rightarrow 6p^3\ {}^2D^o_{3/2}$ transition has been critically evaluated[4] (τ = 190 nsec). Construction of several low-inductance dischrage configurations which produced excitation current pulse risetimes, τ_R, as low as 25 nsec ($\tau_R << \tau$) were also unsuccessful in producing laser action.[2,4]

Williams, Trajmar and Bozinis[5] (WTB) have carried out an experimental determination of electron impact cross sections for excitation of both the upper $^4P_{1/2}$ and lower $^2D_{3/2}$ proposed bismuth laser levels. On the basis of their measurements for 40eV incident electrons Trajmar and Williams suggested[6] that the electron impact cross sections are small, are about the same for upper and lower levels and do not provide a basis for an efficient production of a propulation inversion. With the aid of the computer model[7] developed at Polytechnic for the dynamics of pulsed metal-vapor lasers, we have examined the possibility of laser action at 4722Å in bismuth vapor and find that the WTB measurements actually predict the production of high-power, efficient laser output. We shall now show that the present state of knowledge of electron excitation cross sections is entirely consistent with the generation of strong laser action at 4722Å and provides no basis to account for its absence.

A computer modeling approach is required for an analysis of potential laser action at 4722Å because of the large number of excitation and relaxation channels as well as the large number of bismuth atomic energy levels involved. The model includes rate equations for the population of as many as 15 energy levels of the laser species which may be directly or indicrectly involved in the development of laser action. Also included are rate equations for the population of an additional species which is modeled by two energy levels -- ground state atomic and ionic levels. This species, usually argon, is present to act as a buffer gas and to facilitate electrical breakdown and discharge.

Primary excitation in the model is provided by electron collisions. Relaxation processes are electron de-exciting collisions, radiative relaxation and diffusion to and de-excitation at the walls of the container. The rate constants for transitions between the various energy levels and for ionization depend on the cross sections for these processes and on the electron number density and energy distribution. The electron energy distribution is assumed to have a Maxwell-Boltzmann distribution throughout the discharge pulse, but the average energy changes as determined by an energy balance equation. Since in most cases the cross sections for ionization or for excitation are not known, they have been estimated using the semiclassical formulas of Gryzinski.

The computer model was applied to the situation is bismuth using the lowest 15 atomic energy levels and the ground level of Bi^+ from Moore's tabulation[8] and the transition probabilities of Corliss and Bozman.[9] Radiation trapping was included. Where values of radiation transition probabilities had not been measured, a wall collision relaxation rate of 10^3 sec^{-1} was assumed. Figure 1 shows a computer print-out for the time development of the absolute magnitude of the capacitor voltage (V), the discharge-tube current density (I) and the blue laser output anticipated at 4722Å(*). The computer print-out models the situation where a 2000-pF capacitor charged to 15 kV has suddenly been applied to a discharge tube 2.5 cm diameter with a hot zone 25 cm long containing 10^{15} Bi atoms/cm^3. A circuit inductance of 1μH has been assumed as well as a laser cavity time constant of 3.3 nsec. To facilitate identification of the expected laser output at 4722Å, a solid line has been drawn through the computer generated points (*). The vertical full scale values are 15 kV voltage, 100 amperes/cm^2 current density and 50 kW laser output power. The full scale horizontal value is 100 nanoseconds.

It is evident from Fig. 1 that in spite of the competing processes within the bismuth atom, the computer model predicts that a population inversion will be obtained in the 4722Å transition under conditions which simulate our experimental apparatus. The peak intensity of the expected laser pulse is 45 kW and full width at half maximum intensity ~90 nsec.

Now let us compare the WTB experimentally-determined electron-impact cross sections[5] with the calculated Gryzinski cross sections used thus far in the model. Unfortunately the WTB measurements were carried out only at an electron energy of 40eV, whereas production of a population inversion in a pulsed discharge will be determined by the cross section values at much lower electron energies. Nevertheless a comparison of the Gryzinski calcuation with the WTB measurement may still be illustrative. At 40eV our Gryzinski calculation of the electron excitation cross sections of the proposed upper laser level ($^4P_{1/2}$) was approximately one-half that of

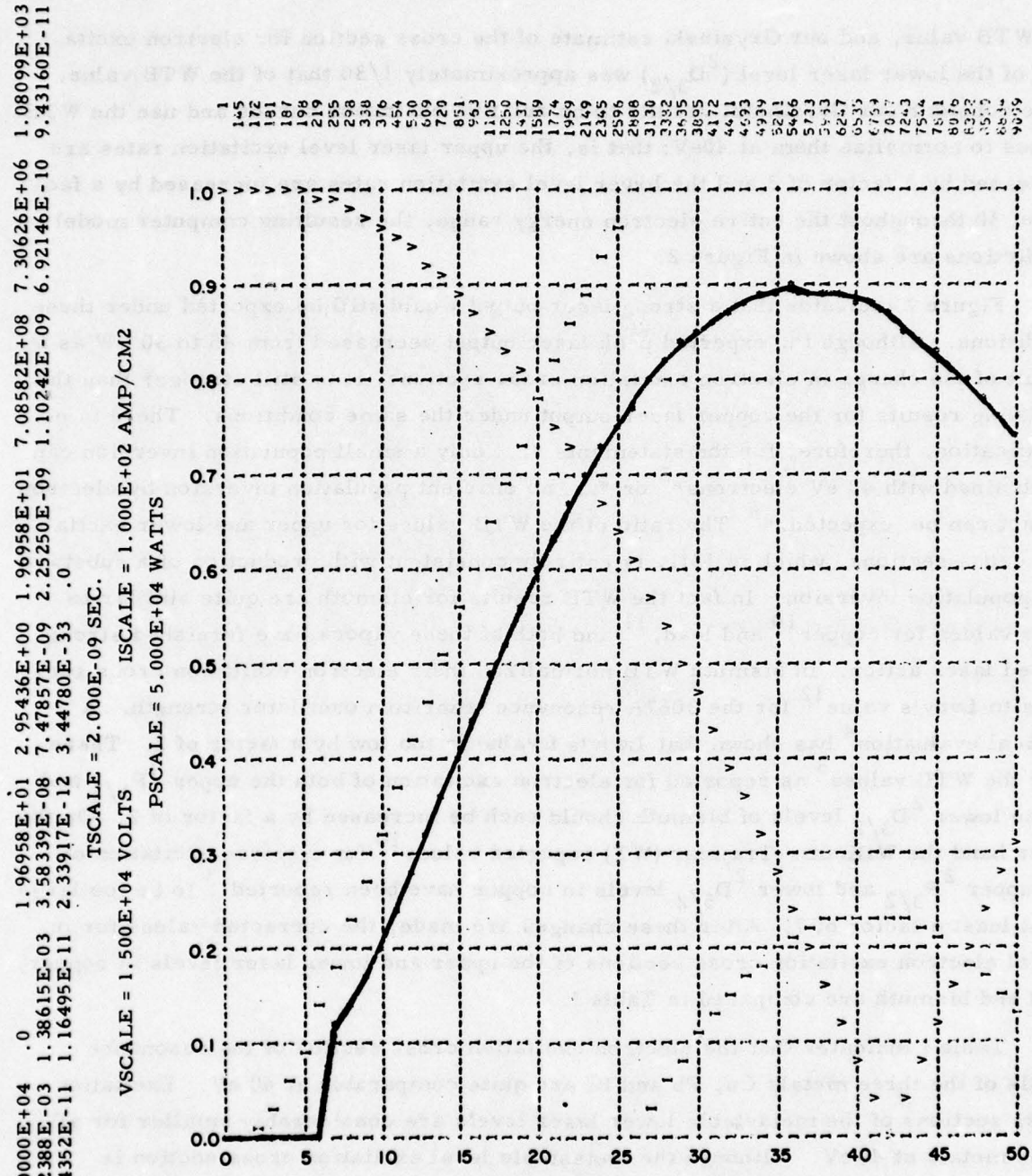


Fig. 1. Computer modeling of a pulsed dischrage in bismuth vapor where the capacitor voltage (V), discharge tube current density (I) and expected 4722Å laser output (*) are shown as a function of time. Gryzinski cross sections are used.

the WTB value, and our Gryzinski estimate of the cross section for electron excitation of the lower laser level ($^2D_{3/2}$) was approximately 1/30 that of the WTB value. If we consider our Gryzinskii results to be relative excitation curves and use the WTB values to normalize them at 40eV; that is, the upper laser level excitation rates are increased by a factor of 2 and the lower level excitation rates are increased by a factor of 30 throughout the entire electron energy range, the resulting computer model predictions are shown in Figure 2.

Figure 2 indicates that a strong laser output would still be expected under these conditions. Although the expected peak laser output decreased from 45 to 30 kW as a result of the change in electron excitation cross sections, it is still stronger than the modeling results for the copper laser output under the same conditions. There is no justification, therefore, for the statements "... only a small population inversion can be obtained with 40 eV electrons"[5] or "... no efficient population inversion by electron impact can be expected."[6] The ratio of the WTB values for upper and lower excitation cross sections, which is 13:1, is entirely consistent with production of a substantial population inversion. In fact the WTB results for bismuth are quite similar to their values for copper[10] and lead,[11] and both of these vapors have furnished strong pulsed laser action. In bismuth WTB normalized their electron excitation cross sections to Lvov's value[12] for the 3067Å resonance transition oscillator strength. A critical evaluation[4] has shown that Lvov's f value is too low by a factor of 2. Therefore the WTB values[5] as reported for electron excitation of both the upper $^4P_{1/2}$ and of the lower $^2D_{3/2}$ levels of bismuth should each be increased by a factor of 2. On the other hand the Williams-Trajmar (WT) reported values[10] for electron excitation of the upper $^2P_{3/2}$ and lower $^2D_{5/2}$ levels in copper have been reported[13] to be too large by at least a factor of 7. After these changes are made, the corrected values for integral electron excitation cross sections of the upper and lower laser levels in copper, lead and bismuth are compared in Table I.

Table I indicates that the electron excitation cross section of the resonance levels of the three metals Cu, Pb and Bi are quite comparable at 40 eV. Excitation cross sections of the metastable lower laser levels are considerably smaller for all three metals at 40 eV. Although the metastable level excitation cross section is larger for Bi than for Cu or Pb, it is still so much smaller than the resonance level excitation cross section that it may be neglected in a first order analysis. For example if the pulsed dischrage consisted entirely of 40 eV electrons, of the Bi atoms excited to one of these two levels 93% would be in the upper laser level.

Again it should be emphasized that it is the electron excitation cross sections at lower electron energies (1 - 10 eV) which will determine the direct production of a

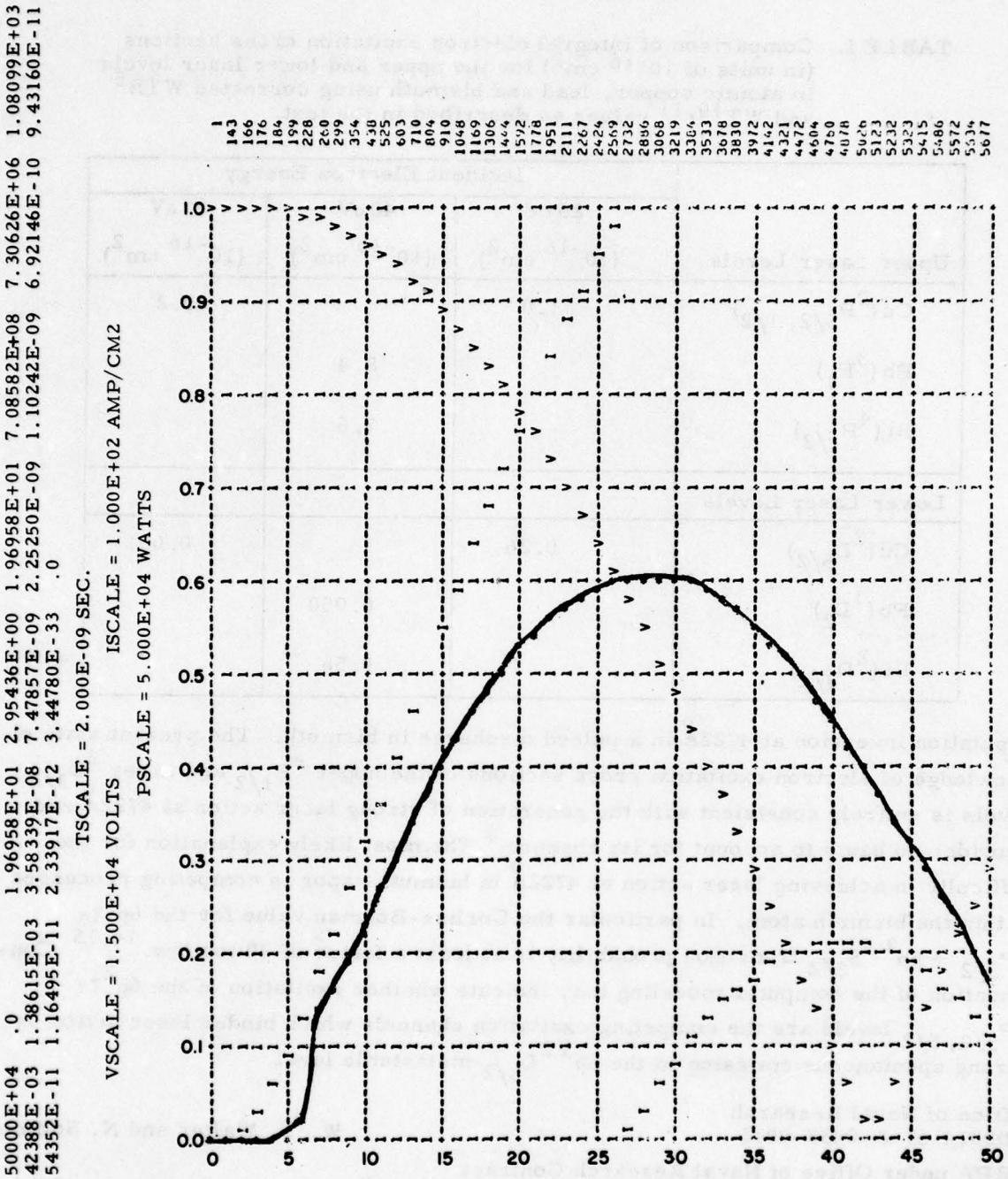


Fig. 2. Computer modeling of a pulsed discharge in bismuth vapor where the capacitor voltage (V), discharge tube current density (I) and expected 4722Å laser output (*) are shown as a function of time. Gryzinski cross sections have been normalized to WTB values at 40 eV.

TABLE I. Comparison of integral electron excitation cross sections (in units of 10^{-16} cm^2) for the upper and lower laser levels in atomic copper, lead and bismuth using corrected WTB[5] and WT[10,11] values as described in the text.

	Incident Electron Energy		
Upper Laser Levels	20 eV (10^{-16} cm^2)	40 eV (10^{-16} cm^2)	60 eV (10^{-16} cm^2)
Cu($^2P_{3/2,\,1/2}$)	11.0		5.2
Pb($^3P_1$)		8.4	
Bi($^4P_{1/2}$)		7.6	
Lower Laser Levels			
Cu($^2D_{5/2}$)	0.26		0.051
Pb($^1D_2$)		0.050	
Bi($^2D_{3/2}$)		0.58	

population inversion at 4722Å in a pulsed discharge in bismuth. The present state of knowledge of electron excitation cross sections of the upper $^4P_{1/2}$ and lower $^2D_{3/2}$ levels is entirely consistent with the generation of strong laser action at 4722Å and provides no basis to account for its absence.‡ The most likely explanation for the difficulty in achieving laser action at 4722Å in bismuth vapor is competing processes within the bismuth atom. In particular the Corliss-Bozman value for the $6p^2 7s\ ^4P_{3/2} \rightarrow 6p^3\ ^4S_{3/2}$ transition probability is at least a factor of 20 too low.[14,15] Continuation of the computer modeling may indicate whether excitation of the $6p^2 7s\ ^4P_{3/2,\,5/2}$ levels are the competing excitation channels which hinder laser action by strong spontaneous emission to the $6p^3\ ^2D_{3/2}$ metastable level.

Office of Naval Research
N00014-67-A-0438-0015

W. T. Walter and N. Solimene

ARPA under Office of Naval Research Contract
N00014-67-A-0438-0017

Joint Services Technical Advisory Committee
F44620-74-C-0056

---

‡After submission of this report, weak stimulated emission at 4722 Å was reported by S. V. Markova, G. G. Petrash and V. M. Cherezov [Sov. J. Quantum Electron., Vol. 7, p. 657 (1977)].

REFERENCES

1. W. T. Walter, N. Solimene, M. Piltch and S. Gould, "Efficient Pulsed Gas Discharge Lasers," IEEE J. Quantum Electron, Vol. QE-2, pp. 474-479 (1966).
2. K. Park, N. Solimene and W. T. Walter, "Investigation of Cyclic Laser Action in Bismuth Vapor," Progress Report No. 38 to JSTAC, Polytech. Inst. of New York, Report No. R-452.38-73, pp. 197-202 (1973).
3. W. T. Walter and K. Park, "Double-Pulse Discharge Examination of Bismuth Vapor for Efficient, Pulsed Laser Action at 4722Å," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76, pp. 120-125 (1976).
4. W. T. Walter and N. Solimene, "Spontaneous-Emission Transition Probability of the 4722Å Atomic Bismuth Line," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76, pp. 126-130 (1976).
5. W. Williams, S. Trajmar and D. G. Bozinis, "Elastic and Inelastic Scattering of 40 eV Electrons from Atomic and Molecular Bismuth," J. Phys. B: Atom. Molec. Phys; Vol. 8, pp. L96-9 (1975).
6. S. Trajmar and W. Williams, "Electron-Metal Atom Collision Cross Sections," (Invited lecture, VIII International Summer School and Symposium on the Physics of Ionized Gases, Dubrovnik, Yugoslavia (1976).
7. N. Solimene and W. T. Walter, "Computer Modeling of the Dynamics of Metal Vapor Lasers," Progress Report No. 38 to JSTAC, Polytech. Inst. of New York, Report No. R-452.38-73, pp. 128-134 (1973).
8. C. E. Moore, "Atomic Energy Levels," National Bureau of Standards Circular 467, Vol. 3, Washington, D.C.: U. S. Government Printing Office (1958).
9. C. H. Corliss and W. R. Bozman, "Experimental Transition Probabilities for Spectral Lines of Seventy Elements," Natl. Bur. Stand. Monograph No. 53, Washington, D.C.: U. S. Government Printing Office (1962).
10. W. Williams and S. Trajmar, "Elastic and Inelastic Electron Scattering at 20 and 60 eV from Atomic Cu," Phys. Rev. Lett, Vol. 33, pp. 187-190 (1974).
11. W. Williams and S. Trajmar, "Elastic and Inelastic Electron Scattering of 40 eV Electrons from Atomic Lead," J. Phys. B: Atom. Molec. Phys. Vol. 8, pp. L50-53 (1975).
12. B. V. Lvov, "Determination of the Absolute Values of Oscillator Strengths by Combined Measurement of the Total and Linear Absorption of Vapor Layers in a Graphite Cell," Optics and Spectroscopy, Vol. 19, p. 282 (1965).
13. D. Trainor, A. Mandl and H. Hyman, "Metal Vapor Visible Laser Kinetics Program," Report No. AD-A022 596/1 WH by AVCO Everett Research Laboratory to ARPA (February 1976). Recently, S. Trajmar, W. Williams and S. K. Srivastava ("Electron Impact Cross Sections for Cu Atoms," submitted to J. Phys. B: Atom. Mol. Phys.) have used helium to calibrate their experimental apparatus and report that their previous cross section values[10] for copper should be reduced by a factor of 2. They estimate an absolute accuracy of a factor of 2 for these reduced values. The helium calibration, however, is involved and depends on several assumptions. On the other hand the calculations of Trainor, Mandl and Hyman are also expected to be accurate to within a factor of 2 and are in agreement with estimates of Cartwright and Leonard. At this present time we shall adopt a factor of 7 reduction of the copper values[10] as the best estimate.

14. W. T. Walter, "A Possible Explanation for the Absence of Laser Action in Bismuth Vapor at 4722Å," Progress Report No. 42 to JSTAC, Polytech. Inst. of New York Report No. R-452.42-77 (1977).

15. W. T. Walter and N. Solimene, "Heat Pipe Bismuth Vapor Laser; Examination of Laser Action at 4722Å in Bismuth Vapor," Report No. Poly EE/EP 76-022 to ONR, Polytech. Inst. of New York (November 1976). NTIS No. AD-A039 568/1WH.

A POSSIBLE EXPLANATION FOR THE ABSENCE* OF LASER ACTION IN BISMUTH VAPOR AT 4722 Å

W. T. Walter

Four possible explanations have been suggested[1] for the absence* of laser action at 4722 Å in pulsed discharges in bismuth vapor:

(1) Insufficiently-fast risetime of the excitation current pulse,[2, 3]

(2) The presence of dimers, Bi_2 molecules, in the vapor,[2, 4]

(3) Unfavorable ratio of electron excitation cross section for the resonance and metastable levels,[5] and

(4) Competing processes; such as, excitation to levels other than the resonance level and cascade filling of the metastable, proposed lower laser level.

In the references listed we have already shown that it is unlikely that any of the first three explanations can account for the absence* of laser action.

Thus far explanation (4), competing processes with the Bi atom also has not provided an answer. The computer modeling results displayed in Figs. 1 and 2 of the preceding article[5] indicate that utilization of Corliss-Bozman transition probabilities[6] to model the rates of competing electron excitation and radiative relaxation processes does not substantially reduce the population inversion expected. The possibility of significant errors in the Corliss-Bozman transition probabilities must be considered. Transition probabilities of some 25,000 classified lines in 70 elements have been calculated by Corliss and Bozman[6] from the intensity tables of Meggers, Corliss and Scribner.[7] Because the Corliss-Bozman values remain the only consistent set of transition probabilities spanning the stronger spectral lines of most of the solid elements, they are extremely useful. Substantial errors, however, have been reported in several of the Corliss-Bozman values, some as large as a factor of 20. Therefore, one is reluctant to rely very strongly on these values. We shall now critically examine these Corliss-Bozman values for bismuth.

As indicated in Fig. 1, the direct laser excitation channel is electron excitation of the 3067 Å $^4S_{3/2} \rightarrow {}^4P_{1/2}$ transition. On the basis of LS coupling, the strengths of competing excitation channels $^4S_{3/2} \rightarrow {}^4P_{3/2 \text{ and } 5/2}$ would be expected to be 2/3 and 1/3 respectively of that of the direct laser excitation channel. The Corliss-Bozman gA values can be converted to Condon-Shortley line strengths[8] by

$$S(J, J') = 4.936 \times 10^5 \lambda^3 gA \quad . \tag{1}$$

The relative Corliss-Bozman line strengths for the $6p^3\,{}^4S_{3/2} \rightarrow 6p^2 7s\,{}^4P_{1/2, 3/2, 5/2}$

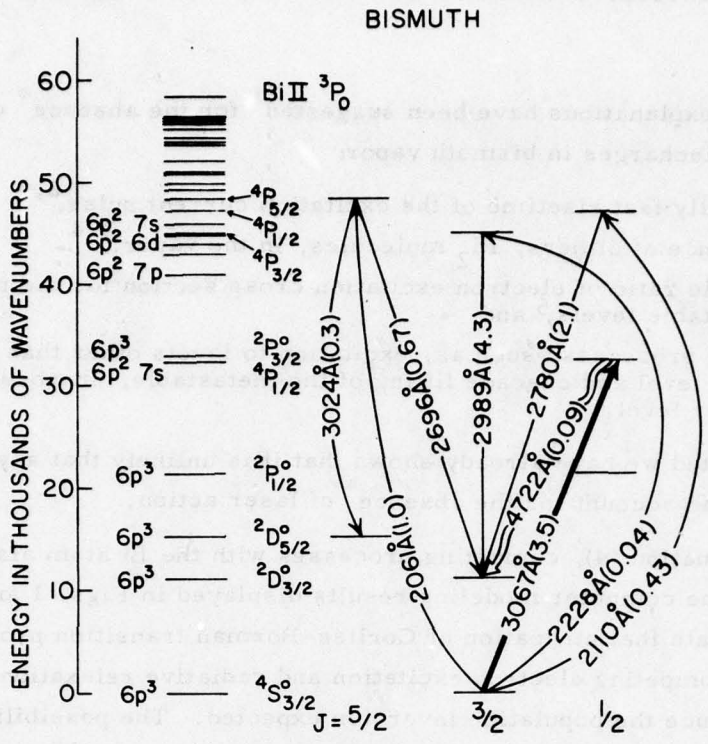


Fig. 1. Proposed pulsed laser transition at 4722 Å in bismuth vapor indicated on a partial energy level diagram of bismuth. Figures in parentheses are Corliss-Bozman[6] transition probabilities in units of 10^8 sec^{-1}.

multiplet then are; 1: .008: .26. Compared with the LS coupling line strength of 1: .667: .333, we see that the strength of the $^4S_{3/2} \rightarrow {}^4P_{3/2}$ transition is 80 times weaker than expected.

When the Corliss-Bozman values for the similar transitions are examined in antimony and arsenic, the elements above bismuth in column VA of the periodic table, the $^4S_{3/2} \rightarrow {}^4P_{3/2}$ member of the multiplet is <u>not</u> observed to be substantially weaker than the other two components. For Sb the Corliss-Bozman relative line strengths for the $5p^3\ ^4S_{3/2} \rightarrow 5p^2\ 6s\ ^4P_{1/2,3/2,5/2}$ multiplet are; 1: 1.39: 3. For As the Corliss-Bozman relative line strengths for the $4p^3\ ^4S_{3/2} \rightarrow 4p^2 5s\ ^4P_{1/2,3/2,5/2}$ multiplet are; 1: .67: (not measured).

Recently intermediate coupling calculations have been carried out for $6p^2 7s \rightarrow 6p^3$ transitions in bismuth.[9,10] The transition probabilities calculated depend on the local exchange approximation as well as on the dipole length or dipole velocity forms of the transition probability operator since the self-consistent field wavefunctions are not exact. The difference between the dipole length and dipole velocity calculations can be

viewed as an indication of the inexactness of the wavefunctions. Line strengths can be calculated from these intermediate coupling transition probabilities by Equation (1). They are listed in Table I for comparison with the Corliss-Bozman line strengths and the LS coupling relative line strengths.

TABLE I. Comparison of line strengths for the resonance multiplets of arsenic, antimony and bismuth. S_{CB} are calculated from Corliss-Bozman's values in Reference 6. S_{DV} and S_{DL} are calculated from Holmgren's values in Ref. 9 for dipole velocity and dipole length approximations. S_{KM} are calculated from Kunisz and Mizdalek's values in Reference 10.

	$^4S_{3/2} \rightarrow {}^4P_{1/2}$	$^4S_{3/2} \rightarrow {}^4P_{3/2}$	$^4S_{3/2} \rightarrow {}^4P_{5/2}$
LS Coupling Relative S	1.000	0.667	0.333
As($4p^3 \rightarrow 4p^2 5s$)			
λ	1972 Å	1937 Å	1890 Å
S_{CB}	1.89	1.26	--
S_{DV}	.848	1.64	2.30
S_{DL}	1.51	2.84	4.00
Sb($5p^3 \rightarrow 5p^2 6s$)			
λ	2312 Å	2176 Å	2069 Å
S_{CB}	.915	1.27	2.75
S_{DV}	1.20	1.95	2.50
S_{DL}	2.45	3.93	4.93
Bi($6p^3 \rightarrow 6p^2 7s$)			
λ	3067 Å	2228 Å	2062 Å
S_{CB}	9.97	0.076	2.60
S_{DV}	3.82	1.55	2.27
S_{DL}	6.75	2.64	3.61
S_{KM}	4.82	4.79	.343

Table I reveals that the Corliss-Bozman line strength for the 2228 Å $^4S_{3/2} \rightarrow {}^4P_{3/2}$ transition in Bi is surprisingly small on the basis of the following comparisons. First, Holmgren's intermediate coupling line strengths for the $^4S_{3/2} \rightarrow {}^4P_{3/2}$ transitions in As, Sb and Bi are comparable (within a factor of 2.5) with the $^4S_{3/2} \rightarrow {}^4P_{1/2}$ line strengths regardless of whether dipole velocity or dipole length approximations are used. Second, the intermediate coupling line strengths of Kunisz and Mizdalek for the $^4S_{3/2} \rightarrow {}^4P_{1/2}$ and $^4S_{3/2} \rightarrow {}^4P_{3/2}$ transitions in Bi are nearly identical. Furthermore, Corliss and Bozman's line strengths for 7 of the 8 transitions

observed in As, Sb and Bi (i.e., all except for the 2228 Å Bi transition) are within a factor of 3 of the intermediate coupling line strengths of Holmgren using either the dipole velocity or dipole length approximations. The Corliss-Bozman line strength of the 2228 Å $^4S_{3/2} \rightarrow {}^4P_{3/2}$ in Bi, however, is at least 20 times smaller than the intermediate coupling line strengths. All indications are, therefore, that this Corliss-Bozman value is anomalously small.

The spectroscopic evidence, both in terms of calculations for Bi as well as calculations and measurements for the analogous elements As and Sb, indicates that the $6p^3\,{}^4S_{3/2} \rightarrow 6p^2 7s\,{}^4P_{3/2}$ transition in Bi is a strong transition. Significant electron excitation of the $6p^2 7s\,{}^4P_{3/2}$ level may be expected, therefore, during a pulsed discharge in bismuth vapor. This expectation is supported by the energy loss spectrum for 40 eV electrons incident upon a bismuth vapor beam, which has been measured by Williams, Trajmar and Bozinis[11] (WTB) and is displayed in Figure 2.

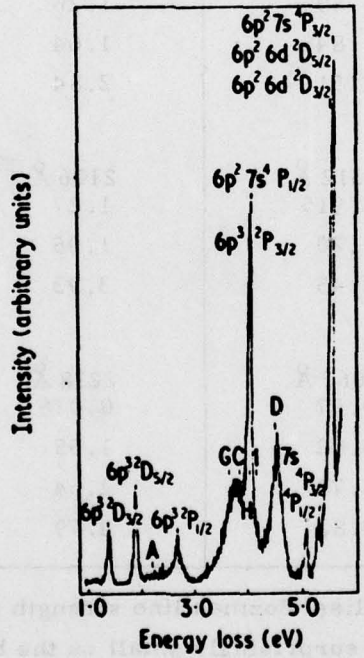


Fig. 2. Energy loss spectrum for 40 eV electrons incident on a bismuth vapor beam measured by Williams, Trajmar and Bozinis. This figure is Fig. 1 in Reference 11.

The first large peak in Fig. 2 must correspond to excitation of the $6p^2 7s\,{}^4P_{1/2}$ resonance level at 4.04 eV and direct excitation of the proposed upper laser level. The other possibility, the $6p^3\,{}^2P_{3/2}$ level at 4.11 eV, is part of the ground level configuration,

and direct excitation should be very small as indicated by the weak strengths of the other $6p^3$ levels at 1.42, 1.91 and 2.69 eV. The second large peak must correspond to excitation of the $6p^2 7s\ ^4P_{3/2}$ level at 5.56 eV. The $6p^2 6d\ ^2D_{3/2,5/2}$ levels at 5.44 and 5.56 eV are not as strongly coupled to the $6p^3\ ^4S_{3/2}$ ground level and are, therefore, not expected to account for the more than a small portion of the second large peak. Finally where the energy loss curve in Fig. 2 ends, it is very likely beginning to rise to a third strong peak at 6.01 eV which corresponds to the $6p^2 7s\ ^4P_{5/2}$ third fine-structure member of the bismuth resonance level.

The WTB energy loss spectrum in Fig. 2 confirms the spectroscopic-based conclusions of preceding paragraphs based on Table I. During a pulsed discharge, therefore, the $6p^2 7s\ ^4P_{3/2}$ level may be expected to be populated at a rate comparable to that of the $^4P_{1/2}$ level. As indicated in Fig. 1 and Table II spontaneous radiation at 2989 Å is expected to be very strong, stronger than the 4722 Å spontaneous radiation, and therefore can seriously interfere with the establishment of the proposed population inversion. Although the several gA values given in Table II for the 2989 Å and 2697 Å lines are not as consistent as they are for the 4722 Å line (variations of 57 and 40 compared with 2), still they suggest that spontaneous radiation of these first two lines will be much more effective in populating the $^2D_{3/2}$ proposed lower laser level than will spontaneous radiation at 4722 Å.

TABLE II. Comparison of spontaneous radiation rates into the $^2D_{3/2}$ proposed lower laser level from the $^4P_{1/2,3/2,5/2}$ resonance levels. gA_{CB} are Corliss-Bozman's values from Reference 6. gA_{DV} and gA_{DL} are Holmgren's values from Ref. 9 for dipole velocity and dipole length approximations. gA_{KM} are Kunisz and Migdalek's values from Reference 10. (The tabulated values are in units of 10^8 photons/sec.)

	$^4P_{1/2} \rightarrow {}^2D_{3/2}$ 4722 Å (10^8/sec)	$^4P_{3/2} \rightarrow {}^2D_{3/2}$ 2989 Å (10^8/sec)	$^4P_{5/2} \rightarrow {}^2D_{3/2}$ 2697 Å (10^8/sec)
gA_{CB}	.18	17.	4.0
gA_{DV}	.132	1.16	.10
gA_{DL}	.099	2.72	.41
gA_{KM}	.091	.30	1.57

The most likely explanation for the absence* of laser action at 4722 Å in bismuth vapor appears to be that the Corliss-Bozman transition probability for the 2228 Å $^4P_{3/2} \rightarrow {}^4S_{3/2}$ Bi line is too low by at least a factor of 20. Strong population of the $6p^2 7s\ ^4P_{3/2}$ resonance level during a pulsed discharge and strong spontaneous radiation at 2989 Å then could fill the $6p^3\ ^2D_{3/2}$ metastable level and prevent the build-up of a sufficient population inversion on the 4722 Å line. Computer modeling studies are continuing to examine this hypothesis. Such computer modeling will also be helpful in evaluating the possibilities of similar potentially-efficient laser transitions in other atomic vapors.

Office of Naval Research
N00014-67-A-0438-0015

W.T. Walter

ARPA under Cffice of Naval Research
N00014-67-A-0438-0017

Joint Services Technical Advisory Committee
F44620-74-C-0056

---

*After submission of this report, weak stimulated emission at 4722 Å was reported by S. Markova, G.G. Petrash and V.M. Cherezov [Sov. J. Quantum. Electron. 7, 657 (1977)]. Stimulated emission was observed only as a thin ring around the o.d. of the discharge tube; it did not fill the tube's cross section. Maximum average-power output was 17 mW which is far less than the 2 W achieved at 6278 Å in gold vapor or 46 W achieved at 5105 Å in copper vapor by Petrash's group in similar discharge-heated apparatus. The explanation presented in this report involving competitive processes in the Bi atom still appears to be the most likely explanation to account for the difficulty in obtaining laser action on the 4722 Å Bi line and for the very weak output power compared with other members of the class of efficient, pulsed atomic-vapor lasers.

REFERENCES

1. W.T. Walter and N. Solimene, "Heat-Pipe Bismuth Vapor Laser; Examination of Laser Action at 4722 Å in Bismuth Vapor," Report No. POLY-EE/EP-76-022 to ONR, Polytech. Inst. of New York (November 1976). NTIS No. AD-A039 568/1WH.

2. K. Park, N. Solimene and W.T. Walter, "Investigation of Cyclic Laser Action in Bismuth Vapor," Progress Report No. 38 to JSTAC, Polytech. Inst. of New York, Report No. R-452.38-73, pp. 197-202 (1973).

3. W.T. Walter and N. Solimene, "Spontaneous-Emission Transition Probability of the 4722 Å Atomic Bismuth Line," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76, pp. 126-130 (1976).

4. W.T. Walter and K. Park, "Double-Pulse Discharge Examination of Bismuth Vapor for Efficient, Pulsed Laser Action at 4722 Å," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76, pp. 120-125 (1976).

5. W.T. Walter and N. Solimene, "Computer Modeling and Examination of Excitation Rates for a Bismuth Vapor Laser at 4722 Å," Progress Report No. 42 to JSTAC, Polytech. Inst. of New York, Report No. R-452.42-77 (1977).

6. C.H. Corliss and W.R. Bozman, "Experimental Transition Probabilities for Spectral Lines of Seventy Elements," Natl. Bur. Stand. Monograph No. 53 (Washington, D.C.: U.S. Government Printing Office, 1962).

7. W.F. Meggers, C.H. Corliss and B.F. Scribner, "Tables of Spectral Line Intensities," Natl. Bur. Stand. Monograph No. 32 (Washington, D.C.: U.S. Government Printing Office, 1961).

8. E.U. Condon and G.H. Shortley, "The Theory of Atomic Spectra," Cambridge University Press (1967).

9. L. Holmgren, "Theoretically Calculated Transition Probabilities and Lifetimes for the First Excited Configuration np^2 (n+1)s in the Neutral As, Sb and Bi Atoms," Physica Scripta, Vol. 11, p. 15 (1975).

10. M.D. Kunisz and J. Migdalek, "Theoretical Values of Certain Transitions of Type $6p^3$-$6p^2 7s$ in the Bismuth Atomic Spectrum in Intermediate Coupling," Acta Physica Polonica, Vol. A47, p. 231 (1975).

11. W. Williams, S. Trajmar and D.G. Bozinis, "Elastic and Inelastic Scattering of 40 eV Electrons from Atomic and Molecular Bismuth," J. Phys. B: Atom. Molec. Phys., Vol. 8, pp. L96-99 (1975).

RATIONAL APPROXIMATION OF TEMPERATURE-DEPENDENT KINETIC COEFFICIENTS FOR EFFICIENT COMPUTER MODELING

G. M. Kull, N. Solimene and W. T. Walter

The modeling of a laser plasma discharge requires the solution of a set of coupled, non-linear, stiff differential equations representing the population densities and temperatures of the plasma components.[1] These equations are coupled through collision processes, such as reaction rates and collision frequencies, which are temperature dependent. Because this system of equations is stiff and non-linear, an iterative implicit method of solution is used. Due to the iterative nature of the solution, the coupling coefficients must be calculated a number of times at each time step. Therefore, an efficient method for representing these coefficients is required.

The first step in efficient approximation is to obtain an approximating function which will have the same behavior as the original function. The most common types of approximating functions are: polynomial series, rational function, trigonometric sum, spline, and special function approximations.[2] The slow speed required for the evaluation of trigonometric and special (i.e. exp, log, etc.) functions eliminates these from prime consideration. Also, the spline method will not be considered, due to its inefficient requirement that the region of interest be subdivided into a set of piecewise polynomial functions. Calculations of the polynomial approximation is efficient, but an inherent problem with this method is its inability to approximate sharp bends within a relatively smooth region. The rational function approximation does not have any of the above problems. Being a ratio of two polynomials, it can be efficiently calculated and can take on the behavior of the function where regular polynomial approximations fail. Thus, the rational function approximation is the logical choice. Determination of the coefficients for the terms of a rational approximation is the difficult part of this method. The procedure which has been developed to optimize the choice of term coefficients to match collision and reaction rates over a broad energy range will now be described.

The collision frequency and reaction rate have been observed to have the following functional dependence on temperature:

$$\langle \nu \rangle \propto f(T^{1/2})$$

$$K(T) \propto g(T^{1/2}) e^{-E^*/kT} .$$

The reaction rate has an additional exponential factor dependent on the activation energy, E^*. This leads to the following choice of a rational function approximation to $f(T^{1/2})$ and $g(T^{1/2})$

$$R(T) = \frac{P(T)}{Q(T)} = \frac{p_0 + p_1 T^{1/2} + p_2 T + p_3 T^{3/2} + \ldots + p_m T^{m/2}}{q_0 + q_1 T^{1/2} + q_2 T + q_3 T^{3/2} + \ldots + q_n T^{n/2}} .$$

The boundary conditions for $f(T^{1/2})$ and $g(T^{1/2})$ are such that as T goes to zero, f and g go to zero, thus $p_0 \equiv 0$. Also, as T goes to infinity f and g also go to zero. Thus the order, n, of the polynomial in the denominator must be greater than the order, m, of the polynomial in the numerator.

As mentioned before, the major problem with the rational approximation is that the p and q coefficients are not straightforward to calculate.[3] The first step in obtaining these coefficients is to determine the ratio of n to m. This is done by obtaining the temperature dependence of f and g as T goes to infinity. Once the ratio of n to m is known, the lowest order of polynomials can be used in P and Q to give the desired fit. This still leaves the problem of obtaining the p and q values. To do this, the maximum error is minimized over a given discrete set of points,[4]

$$r = \underset{i}{MAX} \left| F(T_i) - R(T_i) \right|$$

where F(T) is either $f(T^{1/2})$ or $g(T^{1/2})$. We must also require Q(T) to be greater than zero to prevent a singularity in the region of interest. Because this method is non-linear, an interpolation method can be used. Expanding the error equation about a given iteration point, k, we obtain the following constraints,

$$\left[F(T) + r_k\right] \frac{\delta Q}{Q_k} - \frac{\delta P}{Q_k} + r \geq - r_k(T)$$

$$-\left[F(T) - r_k\right] \frac{\delta Q}{Q_k} + \frac{\delta P}{Q_k} + r \geq r_k(T)$$

where $\delta Q = Q - Q_k$, $\delta P = P - P_k$, and $r_k = F(T_k) - R(T_k)$. Also, for this Taylor expansion to be valid we must add the following constraints,

$$|\delta q_i| < 1 \qquad i = 0, 1, \ldots, n$$

$$|\delta p_i| < 1 \qquad i = 0, 1, \ldots, m \quad .$$

The above set of constraints are solved by linear programming techniques for δq and δp to minimize r. If r is not the best Tchebycheff approximation, the iteration is continued using the new values of P and Q as the old values P_k and Q_k.

A computer program has been written which implements the algorithm to obtain the best rational approximation. As a demonstration of its ability we have obtained the

lowest order rational approximation to the argon momentum-transfer rate. Figure 1 shows a comparison with the experimental collision rate.[4] The agreement is quite

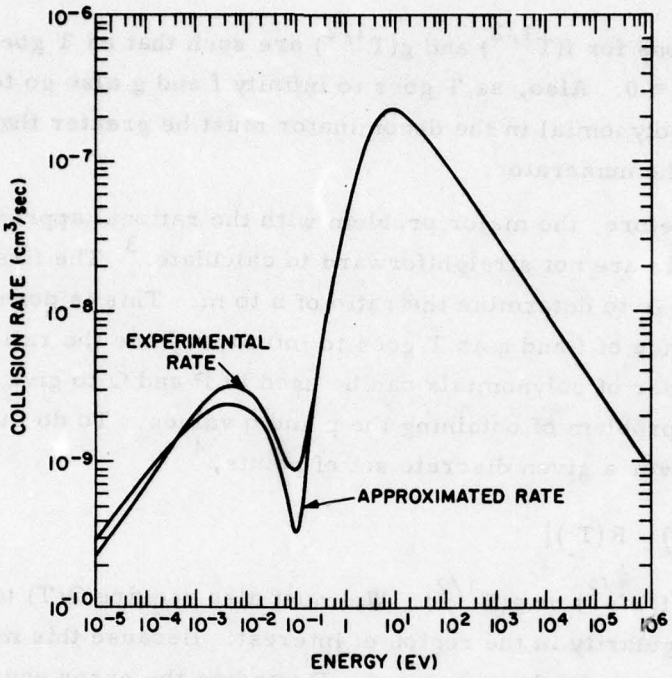


Fig. 1 Argon momentum-transfer rate.

good. Term coefficients (p_i and q_i) for the argon momentum-transfer rate rational approximation are given in Table I.

TABLE I. Term coefficients for rational function approximation of the argon momentum-transfer rate.

$p_0 = 0.0$	$q_0 = 9.379518 \times 10^{-2}$
$p_1 = 9.118165 \times 10^{-9}$	$q_1 = 1.00$
$p_2 = -5.579821 \times 10^{-8}$	$q_2 = -1.375039 \times 10^{-1}$
$p_3 = 8.954906 \times 10^{-8}$	$q_3 = -3.276637 \times 10^{-2}$
	$q_4 = 9.134490 \times 10^{-2}$

Comparison of the computer time required to calculate the momentum-transfer collision rate using the rational approximation with the time taken for the calculation using an integration over the experimental data points or a partial wave representation

indicates that the rational function approximation calculation is about ten times faster. We have shown that the use of rational approximations for reaction rates and collision frequencies is an efficient method of representation over the broad energy region of interest.

Joint Services Technical Advisory Committee
F44620-69-C-0047

G. M. Kull

REFERENCES

1. N. Solimene and W. T. Walter, "Computer Modeling of the Dynamics of Metal Vapor Lasers," Progress Report No. 38 to JSTAC, Polytech. Inst. of New York, Report No. R-452.38-73, pp. 128-133 (1973).
2. John R. Rice, "The Approximation of Functions, Vol. 1, Linear Theory," (Reading, Massachusetts: Addison-Wesley, 1964).
3. E. W. Cheney and T. H. Southard, "A Survey of Methods for Rational Approximation, with Particular Reference to a New Method Based on a Formula of Darboux," SIAM Rev. Vol. 5, pp. 219-31 (1963).
4. I. Barrodale, M. J. D. Powell and F. D. K. Roberts, "The Differential Correction Algorithm for Rational ℓ_∞-Approximation," SIAM J. Numer. Anal. Vol. 9, pp. 493-504 (1972).
5. L. J. Kieffer, "A Compilation of Electron Collision Cross Section Data for Modeling Gas Discharge Lasers," JILA Information Center Report 13, University of Colorado, 13 (1973).

THE CHARACTERIZATION OF NOISE SOURCES BY ANGULAR CORRELATION MEASUREMENTS

D. B. Scarl and S. McAfee

Second order angular field correlation measurements can provide information about the shape and intensity of a distant noise source or scatterer. These measurements, performed by averaging the product of the intensities at two field points, are phase independent and therefore are very little degraded by fluctuations of the medium in which the noise field is propagating. Acoustic and electromagnetic noise sources in the microwave and optical regions can be characterized by second order measurements made by an array of fixed detectors. In contrast to imaging and to first order interferometry which depend on the addition of field amplitudes and are therefore strongly phase sensitive, second order or intensity correlations are sensitive only to propagation delays or fluctuations that are comparable with the correlation time of the detected signal. This correlation time can be, in the case of optical sources, orders of magnitude greater than the period of the propagating waves.

Some of the features of second order correlation measurements can be outlined by considering a simple two-dimensional wave propagation problem. If a source lying along the x axis radiates into the half-plane $z > 0$, the radiation pattern depends on $A(x)$, the source amplitude distribution. A completely coherent source such as an antenna or an acoustic transducer radiates an angular pattern that, in the far field, is the Fourier transform of $A(x)$, with $X = kx$ playing the part of the transform variable t and θ, the angle between the z axis and the field point, playing the part of ω. For example, a source extending from $+\ell$ to $-\ell$ and radiating with uniform amplitude and phase generates an angular amplitude distribution proportional to $\sin(L\theta)/L\theta$. (Here, $L = k\ell$.)

If the same source radiates incoherently with the same intensity distribution as before, the angular intensity distribution in the far field is very different, falling off with $\cos(\theta)$ but providing almost no information about the size or excitation of the source. However, for this coherent source, the first order angular correlation function in the far field is proportional to the Fourier transform of the intensity distribution across the source. For an incoherent source extending from $+\ell$ to $-\ell$ the correlation function is proportional to $\sin(L\Delta\theta)/L\Delta\theta$, where $\Delta\theta$ is the angular separation between detectors. The correlation function, depending only on the angle between the detectors and not on the angle between the detectors and the z axis, gives almost no information about the location of the source, but does give information about its size and intensity distribution. The Fourier transform relationship between source intensity and field angular correlation is independent of phase variations at or near the source, but is affected by phase differences arising from index of refraction fluctuations or motion at or near the

detectors. A characterization of fields in terms of angular correlation functions plays an important part in Wolf's new theory of radiative energy transfer in electromagnetic fields.[1]

If the second order correlation function of an incoherently radiating or scattering source is measured, the part of the correlation function that is dependent on $\Delta\theta$ is proportional to the square of the Fourier transform of the intensity distribution across the source. A linear source extending from $+\ell$ to $-\ell$ would create a second order angular coherence pattern proportional to $1 + S\,(\sin(L\Delta\theta/L\Delta\theta)^2$, where S is a factor having to do with signal to noise ratios and detector properties. Now, however, the signal is independent of phase changes from turbulence or motion either near the source or near the detectors, as long as those changes are less than the coherence time of the signal. The coherence time is set by the source spectrum and is often much greater than the period of the central frequency component. Second order correlation measurements have been used by Hanbury Brown to measure the angular diameter of stars with an intensity interferometer in which the detector separation can be as large as 200 meters.[2] Although the correlation pattern from stars has a very large spatial extent on the surface of the earth, the angle this correlation pattern subtends at the star is less than 10^{-15} radians.

Our recently completed measurement of the second order correlation from a laboratory source of small extent has detected a correlation pattern that extends over an angle of about 0.3 radians as viewed from the source.[3] This measurement was done with two detectors whose large sensitive areas overlapped completely. We have now developed a new brighter chaotic source which will allow us to measure large angle correlations with two small area non-overlapping detectors. These new measurements should show that large angle second order correlation measurements are able to give information about the intensity distribution of sources or scatterers whose size is comparable with the central wavelength of the radiated or scattered noise signal.

Joint Services Technical Advisory Committee
F44620-75-C-0056

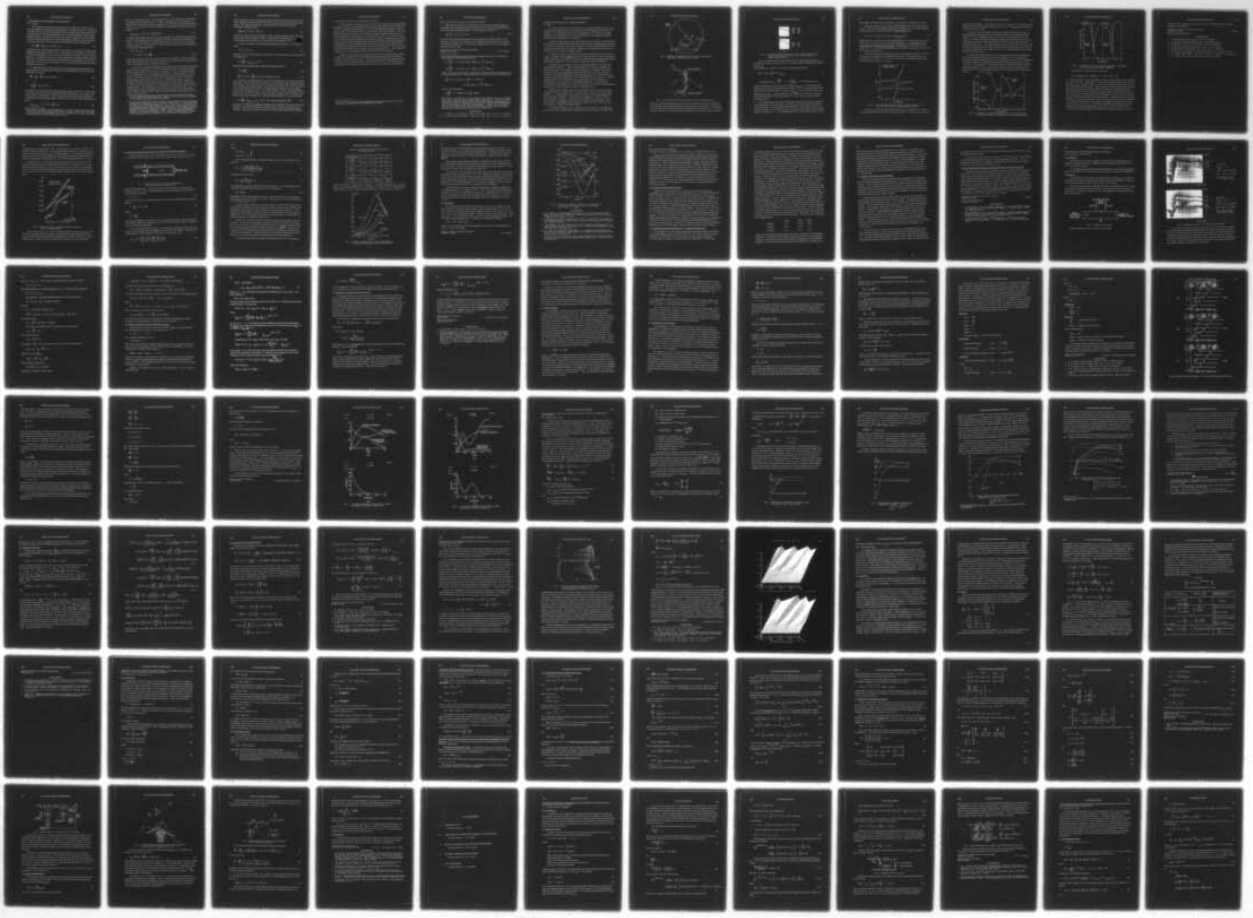
D.B. Scarl and S. McAfee

REFERENCES

1. E. Wolf, "New Theory of Radiative Energy Transfer in Free Electromagnetic Fields," Phys. Rev., D13, pp. 869-886 (February 1976).
2. R. Hanbury Brown, "The Intensity Interferometer," (New York: Wiley, 1974).
3. D. Scarl and S. McAfee, "Two Photon Time Correlations in Diverging Spatially Coherent Light," Proc. 4th Rochester Conference on Coherence and Quantum Optics, Plenum Press (to be published).

AD-A050 265 POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 20/3
PROGRESS REPORT NO. 42, 15 SEPTEMBER 1976 - 14 SEPTEMBER 1977, --ETC(U)
NOV 77 A A OLINER F44620-74-C-0056
UNCLASSIFIED POLY-MRI-452.42-77 NL

3 OF 7
AD
A050 265



A CHALLENGE FOR NONLINEAR SPECTROSCOPY

S. R. Barone

The resolution of saturation spectroscopy is now sufficient to observe the recoil shift between the emission and absorption lines of light molecules.[1] The numerical value of the recoil shift is determined by energy-momentum conservation and is in good agreement with the experimental value. The understanding of future experimental results on, for example, the pressure and power dependence of line shapes, requires a dynamical theory that contains assumptions beyond the conservation laws. One group of theories[2] is partly based on the assumption that the Hamiltonian*

$$H_o = \frac{p^2}{2M} + \frac{1}{2}\hbar\omega_o\,\sigma_3 - \mu E(z,t)\,\sigma_1 \quad , \tag{1}$$

is appropriate for the description of the orbital motion of a molecule in an externally applied, linearly polarized field, $E(z, t)$. The purpose of this report is to point out: 1) the above Hamiltonian contains a force law which has no analog in classical electrodynamics, 2) a modified version of the above Hamiltonian is consistent with classical force laws, 3) experimental results sufficiently precise to establish the correct electromagnetic force on a molecular dipole and the corresponding Hamiltonian would have significant implications.

1) With regard to the first point notice that the Heisenberg equations for the orbital motion of a molecule which follow from the Hamiltonian Eq. (1) and the usual commutation rules are

$$\frac{dz}{dt} = \frac{p}{M} \quad , \quad \frac{dp}{dt} = \nabla_z\,\mu\,\sigma_1\,E(z,t) \quad , \tag{2}$$

or

$$M\,\frac{d^2z}{dt^2} = \nabla_z\,\mu\,\sigma_1\,E(z,t) \quad . \tag{3}$$

According to the usual interpretation of Heisenberg equations the force causing orbital acceleration of the molecule is proportional to the axial gradient of the electric field at the location of the molecule. This is inconsistent with classical electrodynamics. According to classical electrodynamics the force on an electric dipole ($\bar{\mu}$) is (Gaussian units)

$$F_{\text{classical}} = \bar{\mu}\cdot\nabla\bar{E} + \frac{1}{c}\,\frac{d\bar{\mu}}{dt}\times\bar{B} \quad , \tag{4}$$

---

* The notation is standard. The molecular mass, rest frame level separation, dipole moment, position and momentum are M, $\hbar\omega_o$, μ, z and p, respectively. The three-vector ($\sigma_1/2$, $\sigma_2/2$, $\sigma_3/2$) satisfies spin - 1/2 commutation rules. z, p satisfy canonical commutation rules and commute with the σ's.

where c is the speed of light. The second term in the above expression is the analog for a dipole, of the Lorentz force on a charge. In a plane wave, for example, the dipole moment oscillates at the same frequency as the field vectors and the two terms in $\bar{F}_{classical}$ are the same order of magnitude. Using Faraday's Law Eq. (4) may be written

$$\bar{F}_{classical} = \nabla(\bar{\mu} \cdot \bar{E}) + \frac{1}{c} \frac{\partial}{\partial t} (\bar{\mu} \times \bar{B}) \quad . \tag{5}$$

To within velocity dependent forces which are negligible for slowly moving molecules, the above result differs from the right hand side of Eq. (3) by the time derivative of $\bar{\mu} \times \bar{B}/c$. For a molecule in a radiation field the two terms in Eq. (5) are the same order of magnitude. It is instructive to notice further that the interaction Hamiltonian $-\bar{\mu} \cdot \bar{E}$ leads to a translational force

$$\bar{F}_o = \nabla(\bar{\mu} \cdot \bar{E}) \quad ,$$

$$= \bar{\mu} \cdot \nabla \bar{E} - \frac{1}{c} \bar{\mu} \times \frac{\partial \bar{B}}{\partial t} \quad , \tag{6}$$

which again consists of two terms of the same order of magnitude. Details of spectroscopic experiments which are sensitive to the force law and not only the conservation laws* can distinguish between Equation (4) and (16).

2) *The linear momentum* $-\bar{\mu} \times \bar{B}/c$ which accounts for the difference between Eqs. (4) and (6) and is made explicit in Eq. (5) is ordinarily orders of magnitude smaller than a transition photon's worth of linear momentum. Similarly, the dipole energy $-\bar{\mu} \cdot \bar{E}$ is ordinarily orders of magnitude smaller than a transition photon's worth of energy. Nevertheless because of their rapid variation both terms are important in determining the orbital motion of a molecule. Physically the energy $-\bar{\mu} \cdot \bar{E}$ is the energy of the induced electric dipole in the prescribed electric field. Similarly $-\bar{\mu} \times \bar{B}/c$ is the linear momentum to be associated with a stationary electric dipole in a magnetic field. This momentum can be regarded as residing in the field viz. The volume integral of the cross product of the electric field due to the induced dipole and the external magnetic field.** The total translational force on a molecule is the

---

* We are assuming that theories based on (4) or (16) can be elaborated in such a way as to contain the conservation laws. Whether or not this can be accomplished in a reasonable way requires further discussion.

** The physical interpretation of the $\bar{\mu} \times \bar{B}$ term and the analogous term for a magnetic dipole has been discussed briefly in S. Barone, Phys. Rev. D 12, 3363 (1975) and S. Barone and M. Narcowich, Phys. Rev. D, 11, 2880 (1975). These papers contain references to the earlier literature on this subject. H. Haus (private communication) has produced a macroscopic model which verifies (for permant magnetic dipoles) that the term analogous to $-\bar{\mu} \times \bar{B}/c$ is equal to the linear momentum in the field. For details see P. Melman, Ph. D. Thesis, "Orbital Motion of Dipoles in Electromagnetic Fields," Polytech. Inst. of New York (1976) Appendix A.

force required to change its orbital momentum (mass time orbital velocity) plus the force required to change the linear momentum of the field due to the presence of the dipole. This suggests writing the classical equations of motion in the form ($\bar{v} \equiv d\bar{r}/dt$ = orbital velocity)

$$\frac{d}{dt}\left(M\bar{v} - \frac{1}{c}\bar{\mu} \times \bar{B}\right) = \nabla(\bar{\mu} \cdot \bar{E}) \quad . \tag{7}$$

Except for nearly stationary molecules $\bar{\mu} \times \bar{B}/c$ is small compared to $M\bar{v}$. Nevertheless the time rate of change of $\bar{\mu} \times \bar{B}/c$ is comparable to the time rate of change of $M\bar{v}$ and both of these quantities are comparable to the gradient of $\bar{\mu} \cdot \bar{E}$.

In order to construct a quantized, Hamiltonian description it is convenient to define

$$\bar{\pi} \equiv \bar{p} + \frac{1}{c}\bar{\mu} \times \bar{B} \quad , \tag{8}$$

where $\bar{p}$ is the canonical momentum conjugate to $\bar{r}$, the orbital location of the molecule. The Hamiltonian

$$H = \frac{\bar{\pi}^2}{2M} + H_{\text{internal}} - \bar{\mu} \cdot \bar{E} \quad , \tag{9}$$

and the usual commutation rules yield the Heisenberg equations

$$\bar{\pi} = M\frac{d\bar{r}}{dt} \quad , \tag{10}$$

$$\frac{d\bar{\pi}}{dt} = \bar{\mu} \cdot \nabla\bar{E} + \frac{1}{c}\frac{d\bar{\mu}}{dt} \times \bar{B} + \text{terms proportional to } \bar{\pi} \quad , \tag{11}$$

thus according to the Hamiltonian Eq. (9) $\bar{\pi}$ is to be identified as the orbital or kinetic momentum. $\bar{\pi}$ is the molecular mass times its orbital velocity (Eq. (10)) and its time rate of change is the classical force on an electric dipole (Eq. (11)). $\bar{p}$ is the canonical or total momentum (orbital + field). The Hamiltonian consists of internal energy, kinetic energy and dipole energy. The cross-terms $(\bar{p} \cdot \bar{\mu} \times \bar{B} + \bar{\mu} \times \bar{B} \cdot \bar{p})/2Mc$ which appear in Eq. (9) but not in Eq. (1) are the electric analog of spin-orbit terms in the theory of electronic energy levels. The entire Hamiltonian Eq. (9) may be written

$$H = \frac{\bar{p}^2}{2M} + H_{\text{internal}} - \bar{\mu}_m \cdot \bar{B} - \bar{\mu} \cdot \bar{E} + \text{terms proportional to } \bar{B}^2 \quad , \tag{12}$$

where $\bar{\mu}_m \equiv \bar{\mu} \times \bar{\pi}/Mc$ is the magnetic moment to be associated with a moving electric moment. In the present situation the effect of the $\bar{\mu}_m \cdot \bar{B}$ term on energy levels is negligible. This term is necessary however in order to obtain Heisenberg equations

of the same form as the classical equations of motion.

3) The theoretical arguments in favor of Eq. (9) are based on the correspondence principle. But is the correspondence principle sufficiently well verified in this arena that it can be applied without question? One can also question the extent to which the quantum theory really has been made to correspond to the classical theory. The force laws are the same. But there is no generally accepted canonical formalism for classical systems with a dipole moment. The total energy in Eq. (9) consists of internal energy, kinetic energy and electric dipole energy. However on eliminating the velocity in favor of the canonical momentum (as in Eq. (12)) a magnetic dipole energy appears. According to Eq. (9) the magnetic dipole energy is really part of the kinetic energy. This might be interpreted as contrary to classical electrodynamics.* If the force law is accepted as primary the energy appears to be given by Equation (9). If the energy expression Eq. (1) is accepted as primary the force law is non-classical.

It might be argued that a non-classical force law for molecular dipole moments is really not in conflict with the correspondence principle because in the classical limit ($\hbar \to 0$) dipole moments vanish. Molecular dipole moments are quantum mechanical entities and so may be subject to a non-classical force law. Alternatively, quantum effects at the molecular level (proportional to $\hbar$) may contribute to the effective

*There is however an analogy to the appearance of the magnetic moment of a Dirac electron as a consequence of Zitterbewgung of its charge.

force in such a way as to give it a non-classical form.*

Should Eq. (9) turn out to be correct Eq. (8) carries the implication that perpendicular velocity components do not commute and hence cannot be simultaneously determined with unlimited precision. To within terms of order $1/c^2$

$$-\frac{c}{\hbar}\left[\pi_z, \pi_x\right] = \frac{1}{c}\,\bar{\mu}\cdot\nabla B_y \quad . \tag{13}$$

It is well known that perpendicular velocity components of charged systems do not commute. According to Eq. (9) there is an analogous situation for neutral systems with a dipole moment. On the other hand according to Eq. (1), velocity components commute with each other.

Joint Services Technical Advisory Committee
F44620-74-C-0056

S. R. Barone

---

* For a bound state of two oppositely charged, spinless particles, the Heisenberg equations for the individual particles are

$$m_1 \frac{d^2\bar{r}_1}{dt^2} = -\nabla_1 \varphi(|\bar{r}_1 - \bar{r}_2|) + q\left[\bar{E}(\bar{r}_1, t) + \frac{\bar{v}_1}{c} \times \bar{B}(r_1, t)\right]$$

$$m_2 \frac{d^2\bar{r}_2}{dt^2} = -\nabla_2 \varphi(|\bar{r}_1 - \bar{r}_2|) - q\left[\bar{E}(\bar{r}_2, t) + \frac{\bar{v}_2}{c} \times \bar{B}(\bar{r}_2, t)\right]$$

where φ is the static interparticle potential and $\bar{E}$, $\bar{B}$ describe an externally imposed classical, radiation field. The time derivative of the total mechanical momentum is

$$\frac{d^2}{dt^2}(m_1\bar{r}_1 + m_2\bar{r}_2) = q\left[\bar{E}(\bar{r}_1, t) - \bar{E}(\bar{r}_2, t)\right] + q\left[\frac{\bar{v}_1}{c} \times \bar{B}(\bar{r}_1, t) - \frac{\bar{r}_2}{c} \times \bar{B}(\bar{r}_2, t)\right]$$

so that in the dipole limit

$$M\frac{d^2\bar{R}}{dt^2} = \bar{\mu}\cdot\nabla\bar{E}(\bar{R}, t) + \frac{1}{c}\frac{d\bar{\mu}}{dt} \times \bar{B}(\bar{R}, t)$$

where $M \equiv m_1 + m_2$, $\bar{R} \equiv (m_1\bar{r}_1 + m_2\bar{r}_2)/M$ is the usual center of mass coordinate and $\mu \equiv q(\bar{r}_1 - \bar{r}_2)$ is the electric dipole moment of the system. Thus for this model of a quantum system the time derivative of the total mechanical momentum of the system is given by the classical force law of Eq. (4). This result is independent of the number of particles involved in the bound state. The above equation however must be solved subject to the constraint $[\bar{R}, d\bar{R}/dt] = i\hbar/M$. This can introduce quantum effects proportional to $\hbar$ into the time evolution of $\bar{R}$.

REFERENCES

1. J.L. Hall, C.J. Borde and K. Uehara, Phys. Rev. Lett., Vol. 37, p. 1339 (1976).
2. J.H. Shirley and S. Stenholm, J. Phys. A.: Math. Gen., Vol. 10, p. 613 (1977).

A SOLUTION OF THE X-RAY "PHASE PROBLEM"

B. Post

The direct determination of the phases of x-ray reflections from single crystal intensity data constitutes one of the oldest and most challenging problems of x-ray physics. It is of particular interest to crystallographers; its solution could make possible the determination of the crystal structures of all substances from which suitable single crystal data can be obtained.

It has been suspected for some time that coherent interactions among diffracted beams which take place when several sets of planes diffract simultaneously, may provide clues to the solution of the "phase problem." Lipscomb,[1] Fankuchen,[2] Eckstein,[3] Miyake and Kambe,[4] Hart and Lang,[5] and others have investigated the problem from that point of view, with only limited success.

In general, the phases of individual reflections vary with choice of the unit cell origin. The phases of the products of groups of structure factors, the sums of whose indices equal zero, are however, invariant to choice of origin, and their determination has physical significance. Our discussion is limited to such phase products.

A procedure for generating n-beams simultaneous diffraction is illustrated schematically in Figure 1. When reciprocal latice point (rlp) H is brought to its diffracting position on the surface of the Ewald sphere, conventional two-beam diffraction takes place, and diffracted beams are directed to O and H. When the crystal is rotated about OH without disturbing the setting of H, additional rlp's are brought to their diffracting positions and n-beam diffraction ($n > 2$) occurs. Figure 2 shows that under such conditions parallel, overlapping, coherent beams are simultaneously directed to the rlp's in diffracting positions. It is evident that simultaneous n-beam diffraction can provide the necessary conditions for interference among discrete coherent beams in simple, controllable form.

Experimental intensities along two-beam diffraction lines, in the immediate vicinity of three-beam diffraction, are shown in Figs. 3(a) and (b) for triplet phase products of ± 1 respectively. The techniques used to record the photographs have been described previously.[6] The specimen investigated was a single crystal of aluminum oxide (corundum). Its structure factor phases are well known. The indices, phases, and magnitudes of the structure factors involved in the three-beam interactions are listed in the Figure captions.

The distinctive feature of Fig. 3 is that in (a) the intensity is essentially symmetric, and in (b) it is asymmetric about the three-beam direction. We show below

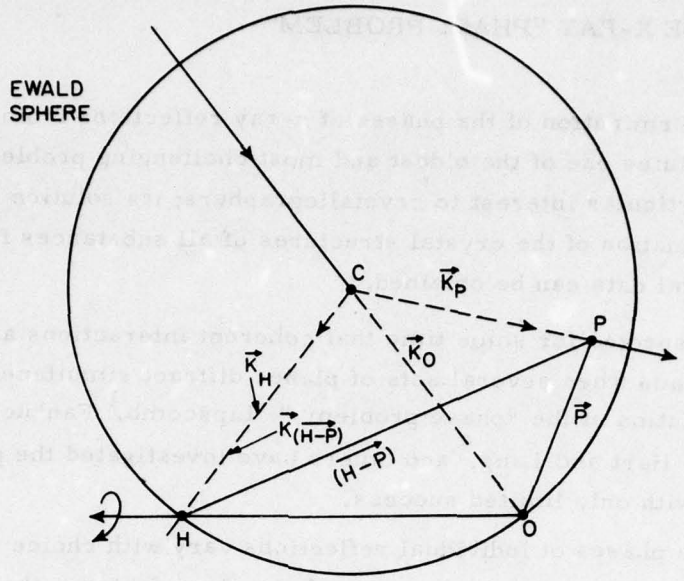


Fig. 1. Schematic representation of three-beam simultaneous diffraction involving rlp's), H, and P.

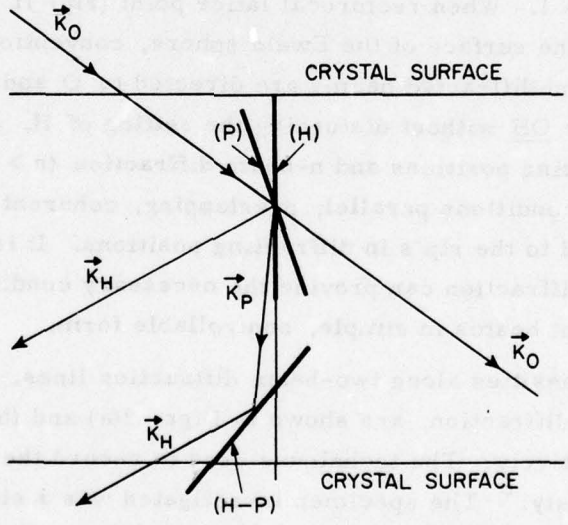


Fig. 2. Discrete diffracted beams directed to rlp H, within the crystal.

that this is a general characteristic of the intensity distributions for the two opposite phase products, provided that all three structure factors are non-zero. The observed symmetry differences are implicit in the dynamical theory of diffraction.[7] They make available a tool for the experimental determination of structure factor phase products.



HKL	F_{CALC}
$1\bar{1}2$	-47
$1\bar{2}0$	+60
012	-47

(a)



$1\bar{1}2$	-47
$01\bar{2}$	-47
104	-81

(b)

Fig. 3. Intensity distributions along two-beam $CuK\alpha_1$, and $K\alpha_2$ lines near three-beam points in Al_2O_3, for: a) positive and b) negative phase products.

The x-ray wave-fields in n-beam diffraction obey Maxwell's equations for a medium with a complex, periodic dielectric constant, under conditions which satisfy Bragg's Law.

The amplitudes are solutions of the linear homogeneous equations:

$$\underline{D}_H \propto (2\epsilon_H)^{-1} \sum_p \phi(H-P)\, \underline{D}_{p[H]} \qquad (1)$$

ϵ_H, defined by $\epsilon_H = \frac{|\underline{K}_H|}{k} - 1 = \frac{\lambda \text{ vac.}}{\lambda \text{ crystal}} - 1$, is an unknown to be chosen such that Eq. (1) has a non-trivial solution. $\phi_{(H-P)}$ is a coefficient in the expansion of the electric susceptibility in a Fourier series. It is proportional to the negative of the structure factor, $(-F_{(H-P)})$. $\underline{D}_{p[H]}$ is the vector component of $\underline{D}_p$ perpendicular to $\underline{K}_H$. The summation is over all rlp's but is limited, in practice, to terms for which the ϵ's are very small, i.e., to the rlp's very near their diffracting positions.

For a three-beam case, Eq. (1) yields three closely similar pairs of solutions. Without significant loss of generality, we will limit our discussion to one member of each such pair. We will discuss centrosymmetric crystals whose origins are at symmetry centers. In such crystals, $\phi_H = \phi_{-H}$ and the phase angles of the structure factors are either 0 or π. It is readily shown that the conclusions reached for three

beams apply, with only minor modifications, to four-beam and higher order cases.

The allowed values of ϵ_H are the solutions of the secular determinant of Equation (1). They are strongly phase dependent, as can be seen from an examination of the expansion of the determinant at the exact three-beam point:

$$(\phi_o - 2\epsilon_o)^3 - (\sum_i |\phi_i|^2)(\phi_o - 2\epsilon_o) + 2\phi_H\phi_P\phi_{(H-P)} = 0 \qquad (2)$$

Equation (2) has 3 unequal real roots summing to zero. The distribution of the signs of the roots is determined by the sign of the last term, i.e., by the sign of the invariant phase of the structure factor triplet. The two possible distributions of the roots (- + + or - - +) lead to propagating modes which differ significantly with respect to excitation and absorption.

The general solution for the ϵ_H near a three-beam point is usually represented by surfaces in reciprocal space defining the wave vector sets ($\underline{K}_o$, $\underline{K}_H$, $\underline{K}_p$), one for each mode, (Figure 4). The collection of surfaces representing all the modes is referred to as the "dispersion surface."

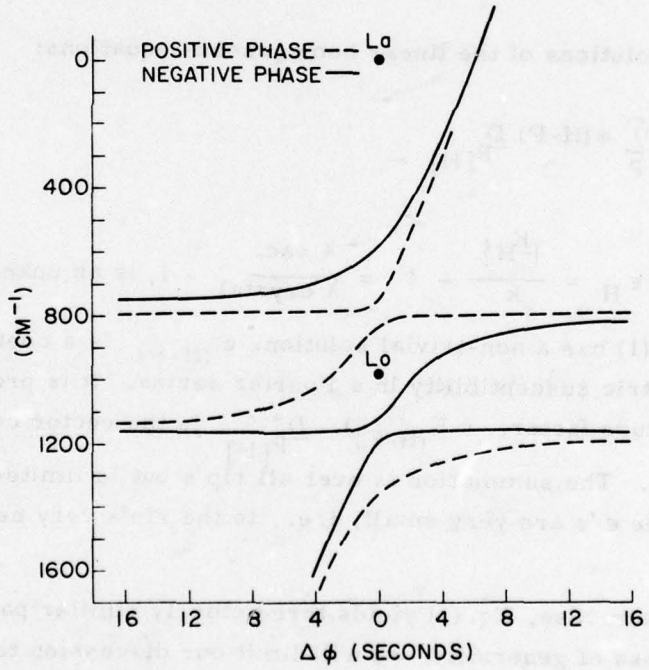


Fig. 4. Section through dispersion surface near three-beam point. Calculated for positive and negative phase products.

A calculated section through the dispersion surface, parallel to a two-beam line passing through a three-beam point near the center, is shown in Fig. 4, for both

signs of the triplet phase. The points La (Laue) and Lo (Lorentz) are the three-beam diffraction points for reciprocal distances of $1/\lambda_{vac.}$ and $1/\lambda_{crystal}$(non-diffracting) from each of the three rlp's.

In general, the modes whose surfaces are closest to the Laue point have the lowest absorption coefficients and are responsible for most of the transmitted intensities.[8] The corresponding modes in Fig. 4 are 1, 2 on the left and 3, 4 on the right. We note that when the triplet phase is positive (dashed lines), the curves corresponding to those modes are approximately symmetrical about the three-beam region (near the center). For negative phase, however, the curves (solid lines) show a large discontinuity just to the right of the three-beam region, and the 3, 4 curve on the right reaches positions equivalent to those of 1, 2 on the left only at large angular distances from the center.

The calculated effects of the above on the excitation of modes of propagation and on the absorption coefficients are shown in Figures 5 and 6. The effects of polarization of the diffracted beams have been taken into account in these calculations. To reduce confusion we show the two-mode averages for only two pairs of modes (1. 2 and 3. 4) in the Figures. For the positive phase product the excitations and the absorption coefficients undergo only minor changes in passing through the three-beam regions. The corresponding changes are much greater for the negative phase product.

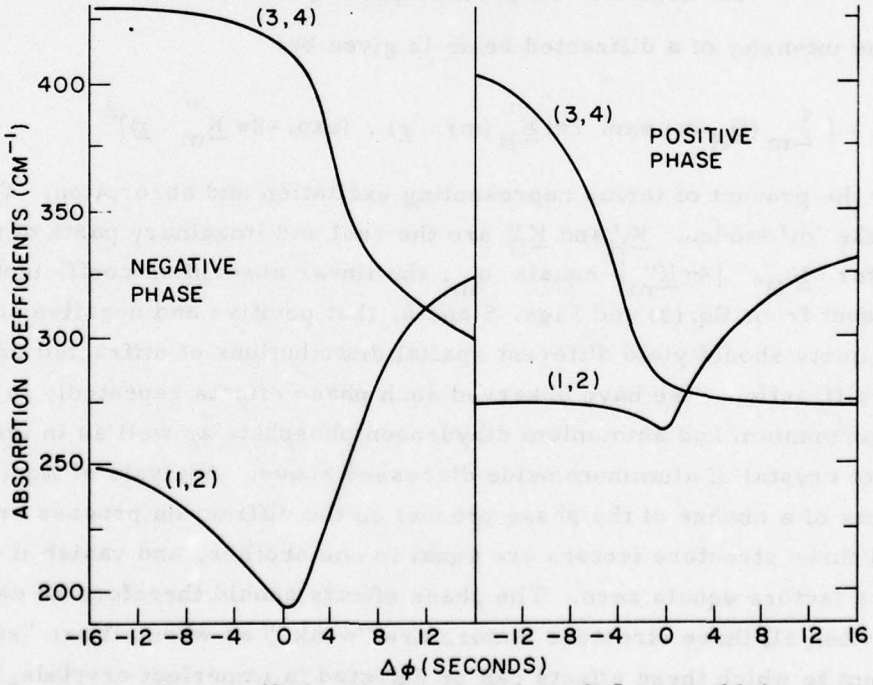


Fig. 5. Absorption coefficients of germanium; 0. 5 mm. thick crystal. Calculated for $CuKa_1$, for negative and positive phase products.

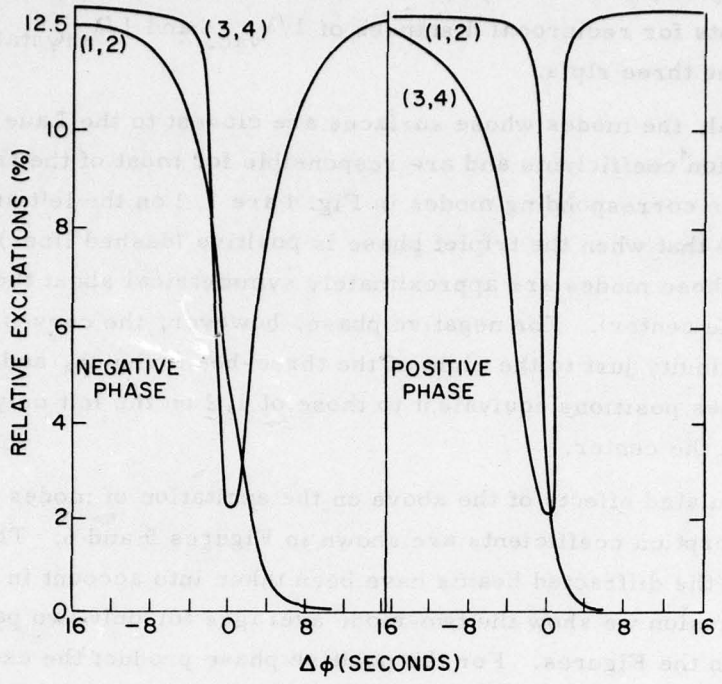


Fig. 6. Relative excitations of modes of propagation. Calculated for negative and positive phase products.

The intensity of a diffracted beam is given by:

$$I_H = [\sum_m (\underline{D}_H(m) \exp.\ 2\pi i\, \underline{K}'_H(m) \cdot \underline{r}) \cdot (\exp. -2\pi\, \underline{K}''_m \cdot \underline{r})]^2 \qquad (3)$$

i.e., by the product of terms representing excitation and absorption. The summation is over the 'm' modes. $\underline{K}'_H$ and $\underline{K}''_H$ are the real and imaginary parts of the propagation vector, $\underline{K}_H$. $|4\pi \underline{K}''_m|$ equals u_m, the linear absorption coefficient of mode 'm.' It is evident from Eq. (3) and Figs. 5 and 6, that positive and negative structure factor phase triplets should yield different spatial distributions of diffracted intensity in n-beam diffraction. We have observed such phase effects repeatedly in perfect crystals of germanium and ammonium dihydrogen phosphate as well as in the relatively imperfect crystal of aluminum oxide discussed above. Analysis of Eq. (2) shows that the effects of a change of the phase product on the diffraction process are maximized when all three structure factors are equal to one another, and vanish if one of the structure factors equals zero. The phase effects should therefore be detected as readily when all three structure factors are "weak," as when all are "strong." The extent to which these effects can be detected in imperfect crystals, such as are

usually used for crystal structure analysis, or in non-centrosymmetric crystals, remains to be determined.

Joint Service Technical Advisory Committee
F 44620-74-C-0056 B. Post
National Science Foundation

REFERENCES

1. W.N. Lipscomb, Acta. Cryst., Vol. 2, p. 192 (1949).
2. I. Fankuchen, Private Communications, quoted in Reference 1.
3. H. Eckstein, Private Communications, quoted in Reference1.
4. S. Miyake and K. Kambe, Acta Cryst., Vol. 7, p. 218 (1954).
5. M. Hart and A.P. Lang, Phys. Rev. Lett., Vol. 7, p. 120 (1961).
6. B. Post, S.L. Chang and T.C. Huang, Acta Cryst., Vol. A33, p. 95 (1977).
7. P.P. Ewald, Ann. Physik, Vol. 49, pp. 1 and 117 (1916); Vol. 54, p. 519 (1917).
8. P.P. Ewald and Y. Heno, Acta Cryst. Vol. A24, p. 5 (1968).

STRAIN DEPENDENCE OF RUST LAYERS AT X-BAND

H. L. Bertoni and L. M. Silber

A. Introduction

Radar returns from moving vehicles, or stationary vehicles with engine running, are found to exhibit modulation, some of which is periodic and some random in nature. Possible sources of this modulation are reflections from moving parts, such as an engine fan, modulation of the radar induced current distribution as a result of intermittent contact between adjacent metal parts, and vibration induced changes in the surface conductivity of rusty metal parts. The nature of radar modulation by agitated metals (RADAM) is of interest as a means of detecting and identifying vehicles, and in order to avoid such detection and identification. Previous investigations have concentrated on the effects of reflections from moving parts,[1] and on the effects of intermittent contacts.[2-5]

The work described here has focused on the importance for the RADAM effect of vibrational strain induced changes in conductivity of rusty surfaces. To examine the effect under controlled conditions, we carried out laboratory measurements at X-band in a waveguide system. In order to enhance the sensitivity of the measurements, a reflection cavity was used, whose end wall was replaced by rusty iron foils. The return loss of the cavity was measured for the foils in both strained and unstrained states. From these measurements the effective surface resistance, the plane wave reflection coefficient, and their dependence on strain, were found.

B. Cavity Measurements

Two techniques were developed to measure the changes in cavity absorption due to straining the rusty foil end-wall of X-band reflection cavities. The first technique was developed to measure the extremely small changes in cavity Q obtained with the lightly rusted foils that were initially prepared. The foils were strained by clamping the edges of the foil to the choke flange at the back end of the cavity and driving the center of the foil in and out via an attached drive pin -- see Figure 1(b). With this method of straining the foil, the effective length, and hence resonant frequency, of the cavity are modulated, as well as the conductivity of the rust film.

The first measurement technique used this resonant frequency modulation. The drive pin was displaced periodically at a frequency f_m and the modulation of the reflected power at frequencies f_m and $2f_m$ was measured. The fixed microwave frequency was first set at the resonant frequency f_o of the unstrained cavity, and then at the half-power frequency $f_{1/2}$. The ratio of modulation at $2f_m$ for microwave frequency f_o to the modulation at f_m for microwave frequency $f_{1/2}$, was determined and compared with

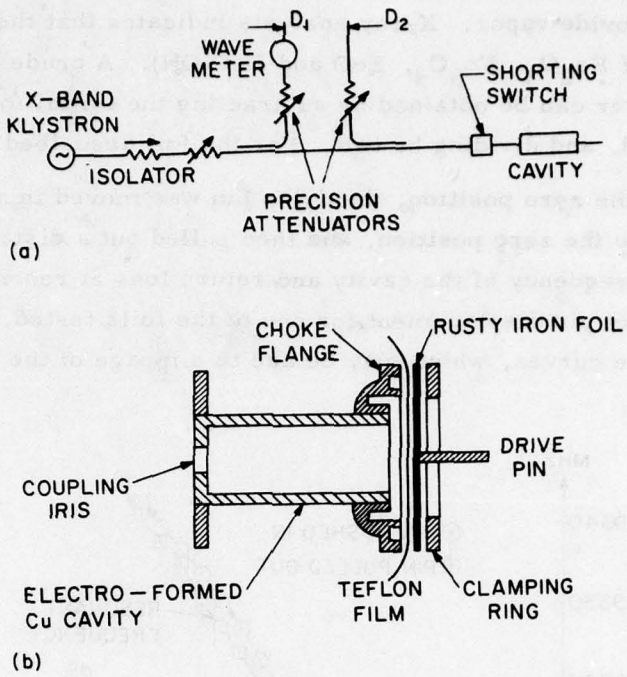


Fig. 1. Set-up for measuring electrical conductivity of rusty iron foils: (a) waveguide circuit; (b) cavity with foil end-wall.

a theory developed assuming no change in conductivity. For copper and clean iron foils the theory and experiment were in good agreement. However, a 30% discrepancy was found between theory and experiment for lightly rusted iron foils, clearly indicating a change in the conductivity of the rust layer. Subsequently, more heavily rusted foils were prepared, for which it was possible to measure the return loss directly, and the modulation technique was not pursued.

Direct return loss measurements were made using a rectangular electro-formed copper cavity. The cavity was approximately one wavelength long at resonance and coupling was via a circular iris, whose diameter was chosen so that the cavity was under-coupled. The measurement circuit is shown in Fig. 1(a) and is driven by a swept klystron oscillator. By keeping the incident power constant and measuring the power reflected from the cavity via a precision attenuator, the reflection coefficient could be measured with a reproducibility of 0.1 dB.

The cavity itself is shown in the cross-section in Fig. 1(b). A 2.5 mills thick teflon sheet was used between the choke flange and foil to eliminate changes in contact resistance as a result of tightening the clamping ring or stressing the foil. A micrometer drive was connected to the drive pin, which was in turn attached to the foil.

Iron foils were rusted by placing 2 mil thick stock in an oven at 70°C, together with a dish of water to provide vapor. X-ray analysis indicates that the rust layer is composed of a mixture of Fe_2O_3, Fe_3O_4, FeO and FeO(OH). A crude estimate of the thickness t of the rust layer can be obtained by subtracting the initial foil thickness from that of the rusted foil, and dividing by two. For the foil described below, t = 1.2 mil.

Starting from the zero position, the drive pin was moved in a distance 0.5 mm and then withdrawn to the zero position, and then pulled out a distance 0.5 mm. The measured resonant frequency of the cavity and return loss at resonance are plotted in Fig. 2 as a function of pin displacement for one of the foils tested. Hysteresis effects are observed in these curves, which may be due to slippage of the foil in the clamping ring.

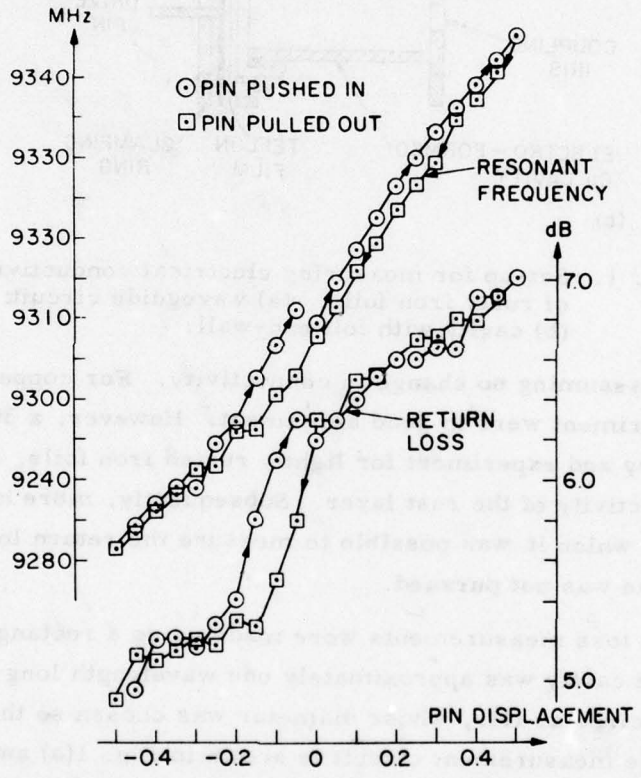


Fig. 2. Measured values of resonant frequency and return loss vs. pin displacement.

While the return loss curve is more irregular than the resonant frequency curve, there is a systematic change with strain. Such a change was <u>not</u> observed for a pure iron foil, and is therefore attributed to the presence of the rust. Because the cavity is undercoupled, increase in return loss indicates a decrease in the loss in the foil.

C. Evaluation of Surface Resistance and Plane Wave Reflection Coefficient

In order to determine the surface resistance of the foil and its plane wave reflection coefficient from the return loss of the cavity, the equivalent circuit of Fig. 3 is

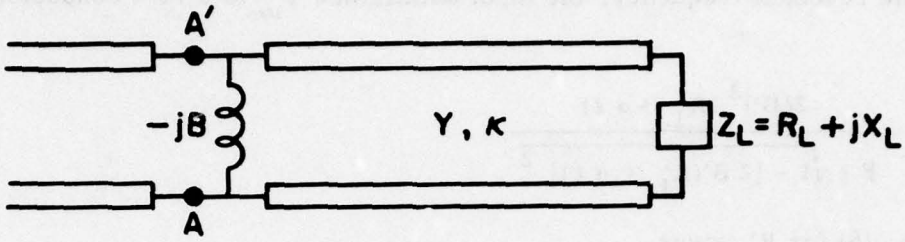


Fig. 3. Equivalent circuit for an iris-coupled cavity with end-wall surface impedance Z_L.

used to describe the cavity. The shunt susceptance -jB with B > 0 represents the coupling iris of the cavity. The transmission line has wavenumber κ, which must be taken complex in the form

$$\kappa = \beta - j\alpha \tag{1}$$

to account for wall loss. The admittance of the line for the dominant mode of the waveguide is

$$Y = \frac{\kappa}{\omega \mu_o} = Y_o\left(1 - j\frac{\alpha}{\beta}\right) \tag{2}$$

where

$$Y_o = \frac{\beta}{\omega \mu_o} , \tag{3}$$

μ_o is the permeability of free space, and ω is the radian frequency. Because the area of the wall containing the coupling iris is only 1/10 that of the side walls, for simplicity the loss in the iris is neglected.

For a conducting end wall $|Y_o Z_L| \ll 1$, while near resonance $\tan \kappa \ell \approx (\kappa \ell - n\pi)$, where n is the number of half-wavelengths nearest to the cavity length. With these approximations, the input impedance to the cavity, normalized to Y_o, is

$$Y'_{in} = -jB' + \frac{(R'_L + \alpha \ell) - j(X'_L + \beta \ell - n\pi)}{(R'_L + \alpha \ell)^2 + (X'_L + \beta \ell - n\pi)^2} , \tag{4}$$

where

$$\left.\begin{aligned} B' &= B/Y_o \\ R'_L + jX'_L &= Y_o Z_L \end{aligned}\right\} \tag{5}$$

At the resonant frequency, the input admittance Y'_{in} is a real conductance G'_{in} given by

$$G'_{in} = \frac{2(B')^2 (R'_L + \alpha \ell)}{1 + \sqrt{1 - [2B'(R'_L + \alpha \ell)]^2}} \quad . \tag{6}$$

Solving Eq. (6) for R'_L gives

$$R'_L = \frac{G'_{in}}{(B')^2 + (G'_{in})^2} - \alpha \ell \quad . \tag{7}$$

The conductance G'_{in} can be found from the return loss ρ, i.e., the magnitude of the reflection coefficient from the cavity load, via the expression

$$G'_{in} = \frac{1 + \rho}{1 - \rho} \tag{8}$$

appropriate for the undercoupled cavity used. Thus, measuring ρ, we can compute R_L knowing the cavity parameters $\alpha \ell$ and B'. The parameters were determined to be $\alpha \ell = 6.12 \times 10^{-4}$ and B' = 43.5.

Using the method outlined above, the normalized surface resistance R'_L and the actual surface resistance R_L were computed for values of return loss ρ nominally corresponding to the zero point and two end points of the pin travel in Figure 2. These values are listed in Table I, together with the normalized input conductance G'_{in} from Equation (8). The remaining data in the table is discussed below. It is seen that the surface resistance R_L changes by a factor of 1.6.

In order to correlate the change in surface resistance with a change in the conductivity σ of the rust it is necessary to postulate an electrical model for rust layer over iron. If we were to assume the rust to be very thick compared to the skin depth for it, then the surface resistance would be given by $R_L = \sqrt{\omega \mu_o / 2\sigma}$. Since R_L decreases as the foil is flexed into the cavity, the thick rust model would imply an increase in the conductivity σ, which is contrary to the expectation that positive strain causes the rust grains to separate slightly thereby decreasing conductivity.

An alternate model that was considered represents the rust as a homogeneous

TABLE I. Measured and computed parameters for rusty iron foil.

Pin Position	in	0	out
ρ(dB)	7	6.2	5
G'_{in}	2.62	2.92	3.56
$R'_L(x10^3)$	0.77	0.93	1.26
R_L	0.412	0.497	0.674
$\sigma(x10^{-3})$	5.9	7.5	11.5
$\lvert\Gamma\rvert$	.9978	.9973	.9964

layer of conductivity σ above the iron substrate. The surface impedance $R_L + jX_L$ and plane wave reflection coefficient were computed as a function of layer conductivity and layer thickness. In Fig. 4 we have plotted the surface resistance R_L as a function σ

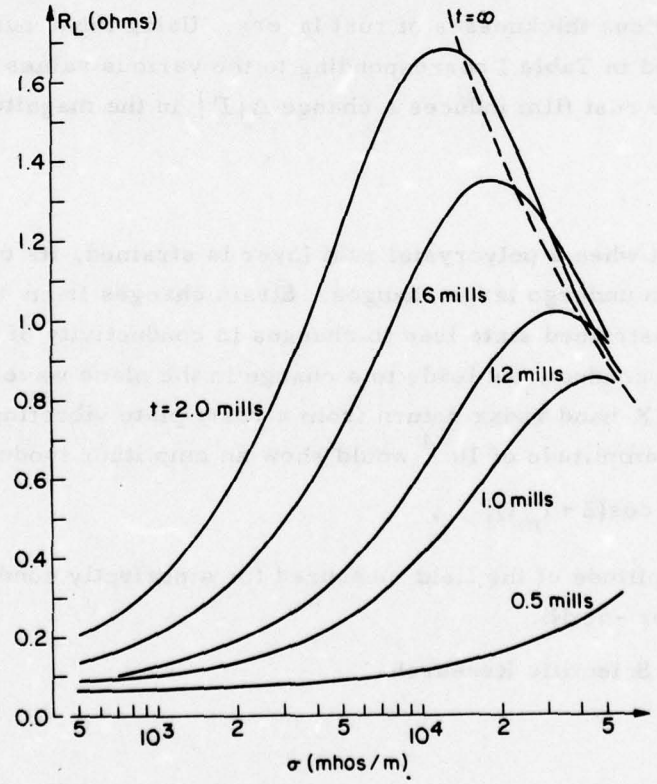


Fig. 4. Surface resistance R_L vs. layer conductivity for rust layers of thickness t over an iron substrate.

for different values of layer thickness t, assuming the frequency to be 10 GHz. It is seen that for finite t, R_L increases with conductivity to a maximum and then becomes asymptotic to the curve for an infinitely thick rust layer. The maximum occurs at a value of σ nearly equal to that for which the thickness t is equal to the skin depth $\delta = \sqrt{2/\omega\mu\sigma}$.

Since single crystal FeO_4 has conductivity 10^4 mhos/m, while Fe_2O_3 and FeO have much smaller conductivity, the effective conductivity of the rust should be less than 10^4 mhos/m. Thus, for rust layers whose thickness is in the range of 1 to 2 mil the surface resistance R_L increases with conductivity.

The values of σ listed in Table I were taken from the curves of Fig. 4 assuming the thickness of the rust layer to be t = 1.2 mil. Because the estimate of layer thickness is very crude, the values of conductivity for each foil may be off by a factor that differs significantly from unity. However, the fractional change in conductivity is not affected by errors in layer thickness. It is seen that σ changes by a factor of almost two as the foil is moved from the "in" position to the "out" position.

The modulation in σ causes a change in the plane wave reflection coefficient Γ. The dependence of Γ on σ is depicted in Fig. 5, where $|\Gamma|$ and arg Γ are plotted as a function of σ for various thicknesses of rust layers. Using these curves, the values of $|\Gamma|$ have been listed in Table I corresponding to the various values of σ. It is seen that the strain in the rust film induces a change $\Delta|\Gamma|$ in the magnitude of Γ that is 1.5×10^{-3}.

D. Conclusion

It is found that when a polycrystal rust layer is strained, its conductivity at X-band frequencies can undergo large changes. Strain changes from $+10^{-4}$ to -10^{-4} on either side of the unstrained state lead to changes in conductivity of almost a factor of two. The change in conductivity leads to a change in the plane wave reflectivity. Thus, since $|\Gamma| \approx 1$, the X-band radar return from a rusty plate vibrating at frequency f_m with induced strain amplitude of 10^{-4} would show an amplitude modulation

$$E_o[1 + \Delta|\Gamma| \cos(2\pi f_m t)] \quad ,$$

where E_o is the amplitude of the field measured for a perfectly conducting plate and $\Delta|\Gamma| = 1.5 \times 10^{-3}$ or -58 dB.

Air Force Office of Scientific Research
AFOSR-76-3030B

H. L. Bertoni

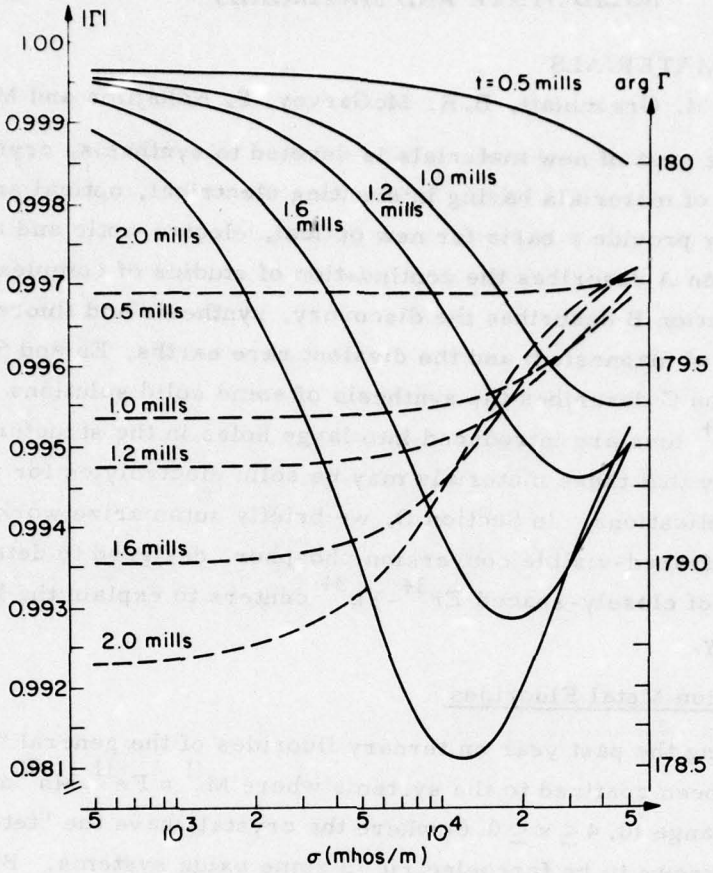


Fig. 5. Plane-wave reflection coefficient vs. conductivity for rust layers of thickness t over an iron substrate: ——— $|\Gamma|$; ----- arg Γ.

REFERENCES

1. R.E. Kleinman, "Electromagnetic Scattering by a Linearly Oscillating Target," Report No. AFCRL-TR-75-0554, Microwave Physics Laboratory, Air Force Cambridge Research Laboratory (1975).

2. A. Hessel and J. Shmoys, "Scattering by a Slit in a Finite Plate and Its Application to the RADAM Problem," Internal Report, Microwave Research Institute, Polytech. Inst. of New York (1975).

3. L.E. Sweeney, Jr., A.J. Bahr and V.R. Frank, "Analysis of Radar Detection of Agitated Metals (RADAM)," Interim Report 1 on AFOSR Contract F44620-75-C-0046, Stanford Research Institute (1976).

4. A.J. Bahr, J.P. Petro and L.E. Sweeney, Jr., "Analysis of Radar Detection of Agitated Metals (RADAM)," Interim Report 2 on AFOSR Contract F44620-75-C-0046, Stanford Research Institute (1977).

5. V.R. Frank, J.P. Petro and A.J. Bahr, "Backscattering from a Cylindrical Dipole Centrally Loaded by a Time-Varying Impedance," IEEE Trans., Vol. AP-25, pp. 356-358 (1977).

NEW SOLID STATE MATERIALS

E. Banks, C. Bush, M. Greenblatt, B.R. McGarvey, S. Nakajima and M. Shone

Research in the area of new materials is devoted to synthesis, crystal growth and characterization of materials having interesting electrical, optical and magnetic properties which may provide a basis for new optical, electro-optic and magneto-optic devices. Section A describes the continuation of studies of complex transition metal fluorides. Section B describes the discovery, synthesis and fluorescence of new mixed fluorides of magnesium and the divalent rare earths, Eu and Sm, $EuMgF_4$ and $SmMgF_4$. Section C describes the synthesis of some solid solutions based on ZrP_2O_7, in which Li^+ ions are introduced into large holes in the structure, and discusses the possibility that these materials may be solid electrolytes for possible advanced battery applications. In Section D, we briefly summarize work on the $CdF_2:Yb^{3+}$, Er^{3+} infrared-visible conversion phosphor, designed to detect the preferential occurrence of closely-spaced Er^{3+}-Yb^{3+} centers to explain the high up-conversion efficiency.

A. Complex Transition Metal Fluorides

Research during the past year on ternary fluorides of the general type $K_xM^{II}_xM^{III}_{1-x}F_3$ has been confined to the systems where $M^{II} = Fe^{II}, Mn^{II}$ and $M^{III}=Fe^{III}$, in the composition range $(0.4 \leq x \leq 0.6)$ where the crystals have the "tetragonal bronze" structure, known to be ferroelectric in some oxide systems. Bridgman pulling of melts yielded boules containing crystals large enough for X-ray analysis. At a nominal composition of $K_{0.5}Mn_{0.5}Fe_{0.5}F_3$, we obtained a single crystal suitable for a structure analysis. Precession photographs showed the unit cell to have its "a" axis equal that of the tetragonal bronze cell, and the "c" axis was doubled. Previously, we had found an orthorhombic superstructure in $K_{0.5}FeF_3$ with $a' = 2\sqrt{2}a$ and $c' = 2c$. The magnetic unit cell found from powder neutron diffraction also showed a doubled "c" axis, but no evidence of a superstructure was seen in the "a" direction. Single crystal intensity data have been collected on $K_{0.5}Mn_{0.5}Fe_{0.5}F_3$ and a structure refinement is now in progress. When the positional and thermal parameters of the atoms are accurately known, we should be in a better position to interpret the neutron diffraction data, and decide whether single crystal neutron data will be required.

B. Fluorescence in New Divalent Rare Earth-Magnesium Fluorides

Several years ago, many papers appeared concerning new ternary fluorides of formula type $BaMF_4$, where $M = Mn^{2+}$, Fe^{2+} Co^{2+}, etc. These materials have an orthorhombic crystal structure, are ferroelectric at and below temperature, and show antiferromagnetic ordering below room temperature. At that time,

we thought it would be of interest to prepare analogous compounds with Eu^{2+} in place of Ba. An attempt to prepare $EuFeF_4$ was made, but we discovered that EuF_2 reduces FeF_2 to iron metal, forming EuF_3. Recently, reports appeared on the non-linear optic behavior of $BaMgF_4$ and $BaZnF_4$,[1] with the same crystal structure as $BaMnF_4$, etc. Since magnesium is reduced to the metal with much more difficulty than the transition metals, we decided to attempt the preparation of $EuMgF_4$. Initial attempts in Pt tubes led to the appearance of a few new weak lines and a blue fluorescence under *U. V. excitation*, characteristic of Eu^{2+}. The platinum was blackened, indicating alloying of Mg. This led us to attempt the use of carbon crucibles. Mixtures of EuF_2 and MgF_2 heated at temperatures below 900° showed very little sign of reaction. Heating to 1100° produced melting and the melt showed only starting materials. This suggested an incongruently melting compound, and very slow solid state kinetics. It was then decided to attempt the reaction via Mg vapor. We mixed stoichiometric amounts of EuF_3, MgF_2 and Mg metal and heated the mixture in a sealed graphite crucible at 800-900°C for several hours. The product showed spotty sites with blue fluorescence. Thus far, the most successful preparations were made by adding Mg metal to the top of a mixture of EuF_3 and MgF_2. The product after heating was a single phase with a new X-ray pattern whose major lines can be indexed on an orthorhombic cell. This cell is comparable to the barium compounds, with smaller dimensions. It has a brilliant blue fluorescence under ultraviolet excitation. Such high brightness at room temperature in a stoichiometric compound is quite unusual, as interaction among the activator ions usually quenches luminescence. We have also prepared $SrMgF_4$, whose unit cell is almost identical with that of $EuMgF_4$, as expected from the similarity in ionic radii of Eu^{2+} and Sr^{2+}. This also appears to be incongruently melting, as a melted sample yielded the patterns of the starting materials. Theother successful preparation using the magnesium vapor method was $SmMgF_4$, with a unit cell similar to that of $EuMgF_4$. The unit cells are compared to that of $BaMgF_4$ below:

	a (A)	b (A)	c (A)
$BaMgF_4$	5. 81	14. 509	4. 125
$SmMgF_4$	5. 88	14. 48	4. 19
$EuMgF_4$	5. 82	14. 39	4. 14

The Sm and Eu compound cell parameters decrease as expected for the lanthanide contraction. The fact that some of the Ba compound parameters are unexpectedly smaller is not understood; it may indicate that the compounds are not actually isostructural. We are now preparing the solid solution samples $Sr_{1-x}Eu_xMgF_4$ for a

study of the fluorescence and excitation spectra as a function of composition. The bright blue emission of these materials suggests that they may be useful X-ray detectors for applications connected with films or photosensitive detectors. We are now attempting the preparation of thin films of $EuMgF_4$ by vacuum evaporation, as such films can be used in evacuated image tubes. We are also attempting to grow single crystals from off-stoichiometric melts. The broad, intense luminescence in this compound suggests the possibility that it may be a material for a solid-state laser which could be tunable over a portion of its emission spectrum, and if it proves to be ferroelectric, the possibility of modulating its luminescence by electric fields may be worth exploring.

C. Synthesis of Potential Solid Electrolytes

Solid electrolytes have become important in the development of rechargeable batteries of high energy density for applications to electric vehicles and utility load-leveling. The most advanced systems of this sort use a sodium-ion conductor based on $NaAl_{11}O_{17}$ (β-alumina) which has two-dimensional ion conduction pathways and is inherently a poorer electrolyte than a material which might have three-dimensional connections among the cation sites, which would be partially occupied to permit free movement of ions from one site to the others.

We have attempted to introduce alkali metal ions into the large interstices in cubic ZrP_2O_7 in the hope that these ions would be mobile. The large interstices can accommodate ions of radius up to about 1.2 A, and they are connected three-dimensionally. The first preparations were made using Y^{3+} as the charge-compensating species. This resulted in appreciable solid solution formation only in the case of lithium. Preparations of composition $Zr_{1-x}Y_xLi_xP_2O_7$ showed a cubic phase for values of x up to 0.1. Higher doping levels showed no further increase in lattice parameter and the appearance of a second phase pattern in X-ray diffractograms. No solid solution was observed for combinations of yttrium-sodium and gallium-lithium. A sample of $Zr_{0.9}Y_{0.1}Li_{0.1}P_2O_7$ has been submitted for ^7Li NMR to determine whether the lithium ions are mobile at temperatures up to 200°C. If mobility is detected, we plan to evaluate the conductivity on the basis of dielectric loss measurements.

On the basis of ionic radius considerations, it was thought that these preparations might have lithium on normal Zr sites and yttrium in the large interstices. The X-ray intensities which are about the same for pure ZrP_2O_7 and the solid solution, discourage such an interpretation. With Y in the Zr sites, the 10% filling of the interstices by lithium would lead to a maximum of 2% change in intensity for the

most sensitive reflections, whereas the distribution suggested above would give easily observable intensity changes.

Trivalent indium has a radius much closer to Zr^{4+} than does Y^{3+}. We are planning similar preparations with this ion. Some preparations with Eu^{3+} - Li^{+} substitution are also planned. The photoluminescence of Eu^{3+} can be used as a probe of the Eu environment, as shown by Blasse, et al.[2]

D. Rare Earth-Defect Complexes and Infrared-Visible Upconversion in CdF_2

Our paper on the infrared-visible light conversion in CdF_2: Yb^{3+} has been published in the Journal of the Electrochemical Society.[3] Measurements of excitation spectra were unsuccessful because of the presence of second order radiation from the source, which directly excited the Er^{3+} radiation, and the low intensity of infrared radiation in the spectrometer source. We have recently obtained evidence from ^{19}F NMR studies on a crystal of CdF_2, 5% YbF_3, that approximately 50% of the Er^{3+} and Yb^{3+} ions are involved in mixed (Er^{3+}, Yb^{3+}, $2F_i^-$) dimers. This was in accordance with our proposal to account for the unusually high upconversion efficiency in the phosphor crystals. In that case, the concentration of Er^{3+} ions is only 10% that of Yb^{3+}, so almost all of the Er^{3+} ions might be in such clusters.

Joint Services Technical Advisory Committee
F44620-74-C-0056

E. Banks

U.S. Army Research Office
DAAG-29-75-G-0096

REFERENCES

1. J.G. Bergman, G.R. Crane and H. Guggenheim, "Linear and Nonlinear Optical Properties of Ferroelectric $BaMgF_4$ and $BaZnF_4$," J. Applied Phys., Vol 46, pp., 4645-4646 (1975).

2. H.E. Hoefdraad, F.M.A. Stegers and G. Blasse, "Evidence for the Influence of an Effective Charge on the Position of the Charge-Transfer of Eu^{3+} in Solids," Chemical Physics Letters, Vol 32, pp. 216-217 (1975).

3. M. Greenblatt and E. Banks, "CdF_2-YbF_3-ErF_3 - An Efficient Infrared to Visible Upconverting Systems," J. Electrochem. Soc., Vol 124, pp. 409-413 (1977).

LARGE SIGNAL TRANSIENT RESPONSE OF MOSFETS

R. T. Kinasewitz and B. Senitzky

A. Introduction

Most studies of the current-voltage characteristics of the metal-oxide-semiconductor field-effect transistor (MOSFET) have been directed towards its D. C. and A. C. (small signal) behavior.

Our investigation is directed towards identifying some of the physical phenomena that are responsible for the large-signal turn-on transient response of the MOSFET.

B. Research

In the last report we described our experimental observation of an unexplained overshoot in the transient response of a p-channel enhancement MOSFET. Since then we have experimentally probed for the cause(s) of the overshoot.

One series of experiments consisted of exciting the MOSFET with a D. C. biased gate-to-source signal to determine the effect on the output transient of an input D. C. bias voltage. The bias circuit shown in Fig. 1 was installed in a General Radio Type 874-x insertion unit.

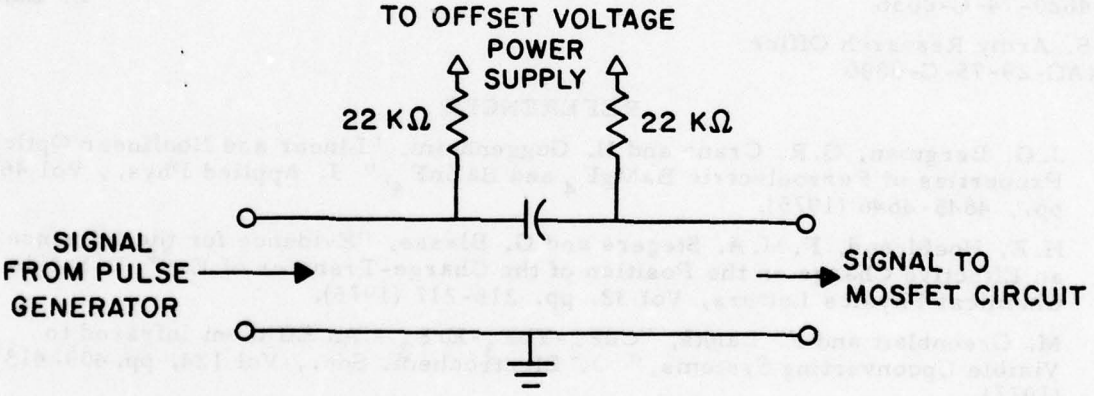
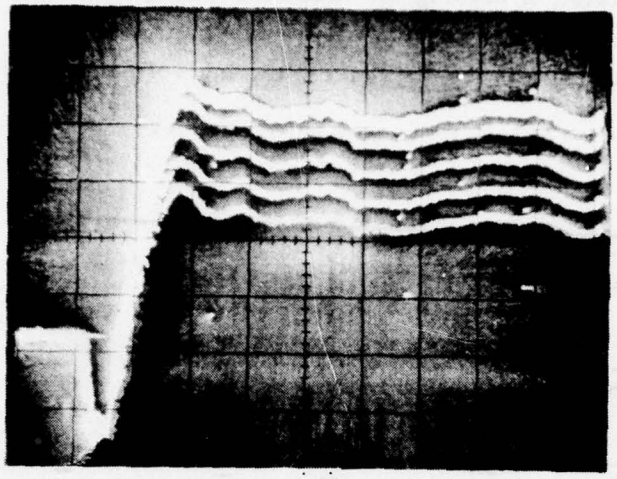


Fig. 1. Offset voltage circuit.

Typical results are shown in Figures 2(a) and 2(b).



V_{OFF}

= -4V, -5V
= -3V
= -2V
= -1V
= -0V

V_{DD} = 20V

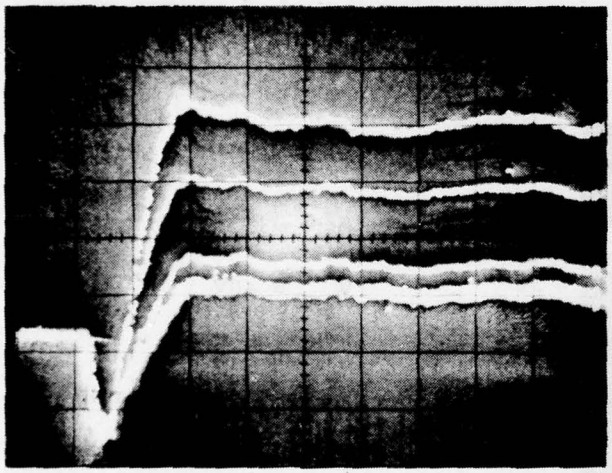
V_{input} = -12V

Vert. scale = 10 ma/Div

Hor. scale = 1 nsec/Div

D.C. gate-to-source bias voltage = V_{OFF}

Figure 2(a).



V_{OFF}

= -6V
= -7V
= -8V
= -9V
= -10V

V_{DD} = 20V

V_{input} = -12V

Vert. scale = 10 mA/Div

Hor. scale = 1 nsec/Div

D.C. gate-to-source bais voltage = V_{OFF}

Figure 2(b).

Fig. 2. MOSFET output vs. D.C. gate-to-source voltage.

In Fig. 2(a) and 2(b) it is seen that a gate-to-source bias voltage more negative than -6V practically eliminates the positive-going overshoot thus showing that this overshoot is an intrinsic device characteristic dependent on the input voltage and not merely an extrinsic capacitance-inductance effect. Extrinsic capacitances and inductances are essentially linear parameters whose transient response is unaffected by D.C. inputs; however, in this experiment the D.C. level of the input signal affected the transient response waveshape very much indeed.

Another series of experiments consisted of biasing the MOSFET with a D.C. substrate-to-source voltage (V_{BS}) to determine its effect on the output transient. The MOSFET circuit was modified to contain the substrate-to-source biasing voltage and is shown in Figure 3.

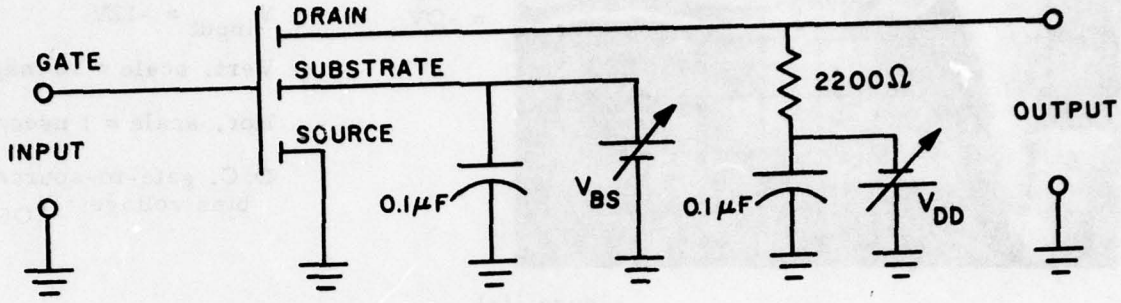


Fig. 3. MOSFET circuit with D.C. substrate-to-source bias.

Typical results are shown in Figure 4.

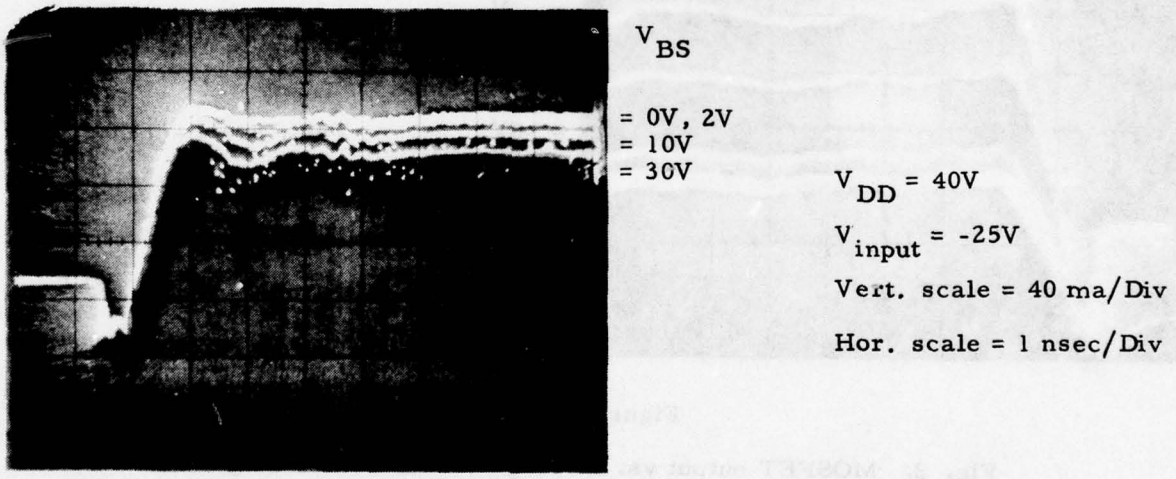


Fig. 4. MOSFET output vs. D.C. substrate-to-source bias.

Figure 4 shows that a substrate-to-source D.C. voltage that back biases the source-substrate P-N junction will cause a decaying oscillation of the output transient.

C. Future Work

Our studies on the large signal subnanosecond transient behavior of MOSFET devices is continuing and we hope that we will be able to elucidate some of the observed behavior in a future report.

Joint Services Technical Advisory Committee
F44620-74-C-0056

R. T. Kinasewitz and B. Senitzky

TURBULENCE IN PLASMA-LIKE SYSTEMS

S. Barone and N. Marcuvitz

An unobjectionable self-consistent description of turbulent electromagnetic fields in plasma-like systems, wherein charges and currents are nonlinearly field dependent is not known. Renormalization procedures applicable in a linear and quasi-linear range fail with the onset of marked nonlinearities. An initial approach to these difficulties is being explored via renormalization of higher point correlation of the electromagnetic fields. For simplicity of illustration we consider an electron plasma.

A. Two-Point Distribution Functions for the Electromagnetic Field

Consider, for simplicity, an electron plasma (charge -e, mass m) with a fixed neutralizing background. Maxwell's equation may be written[1]

$$Y_M \cdot \hat{\underset{\sim}{E}} = e \int (d\underset{\sim}{v})\, \underset{\sim}{v}\, \hat{f}(\underset{\sim}{v}) \quad ,$$

where the dyadic operator Y_M is the same as in I and $\hat{f}$ is the stochastic electron distribution function. According to Eq. (37a) of Ref. (1)

$$\hat{f} = \hat{f}^a + G\, \tilde{T}_c\, \hat{f}^a \quad ,$$

where we have defined $\hat{f}^a \equiv G\hat{\eta}$. If $\tilde{T}_c$ is approximated by

$$\tilde{V} \cong \frac{e}{m}\, \tilde{\underset{\sim}{E}} \cdot \nabla_{\underset{\sim}{v}} \quad ,$$

Maxwell's equations take the form

$$(Y_M - \hat{V}_e) \cdot \hat{\underset{\sim}{E}} \cong -\hat{\underset{\sim}{J}} \quad ,$$

where

$$\hat{V}_e \equiv \frac{e^2}{m} \int (d\underset{\sim}{v})\, \underset{\sim}{v}\, G(\underset{\sim}{v}, \underset{\sim}{v}')(d\underset{\sim}{v}')\, \nabla_{\underset{\sim}{v}'}\, \hat{f}^a(\underset{\sim}{v}') \quad ,$$

$$\hat{\underset{\sim}{J}} \equiv (-e) \int (d\underset{\sim}{v})\, \underset{\sim}{v}\, \hat{f}^a(\underset{\sim}{v}) + \hat{V}_e \cdot \underset{\sim}{E} \quad .$$

These equations have the same structure as the Klimontovich equation so that the techniques of Ref. (1) apply directly.

Following Ref. (1) we introduce a stochastic field Green's function $\hat{Z}$ by

$$(Y_o - \hat{V}_e) \cdot \hat{Z} = 1 \quad ,$$

where $Y_o \equiv Y_M - V_e$. The ensemble average field Green's function satisfies

$$(Y_o - V_{ce}) \cdot Z = 1 \quad ,$$

which might be taken as a defining equation for V_{ce}. A $\tilde{T}_c$ operator is defined by

$$\hat{Z} = Z + Z \cdot \tilde{T}_{ce} \cdot Z \quad .$$

The stochastic, two point distribution function for the field satisfies

$$(Y_o - \tilde{V}_e)_1 (Y_o - \tilde{V}_e)_2 : \hat{\underline{E}}(1)\hat{\underline{E}}(2) = \hat{\underline{J}}(1)\hat{\underline{J}}(2) \quad .$$

Defining

$$Y_{12} = Y_o(1)\, Y_o(2) + \langle \tilde{V}_e(1)\, \tilde{V}_e(2) \rangle \quad ,$$

$$\tilde{V}_e^{(2)} = Y_o(1)\, \tilde{V}_e(2) + \tilde{V}_e(1)\, Y_o(2) + \langle \tilde{V}_e(1)\, \tilde{V}_e(2) \rangle - \tilde{V}_e(1)\, \tilde{V}_e(2) \quad ,$$

this may be written

$$(Y_{12} - \tilde{V}_e^{(2)}) : \hat{\underline{E}}(1)\, \hat{\underline{E}}(2) = \hat{\underline{J}}(1)\, \hat{\underline{J}}(2) \quad .$$

A two-particle stochastic Green's function may be introduced by

$$(Y_{12} - \tilde{V}_e^{(2)}) : \hat{Z}_{12} = 1 \quad .$$

The average satisfies

$$(Y_{12} - V_{ce}^{(2)}) : Z_{12} = 1 \quad ,$$

which defines $V_{ce}^{(2)}$. Also a two-particle version of $\tilde{T}_{ce}$ may be defined by

$$\hat{Z}_{12} = Z_{12} + Z_{12} : \tilde{T}_{ce}^{(2)} : Z_{12} \quad .$$

$V_{ce}^{(2)}$ is related to $\tilde{T}_{ce}^{(2)}$ by

$$V_{ce}^{(2)} = \langle \tilde{V}_e^{(2)} : Z_{12} : \tilde{T}_{ce}^{(2)} \rangle \quad .$$

The solution for $\hat{\underline{E}}(1)\, \hat{\underline{E}}(2)$ is

$$\hat{\underline{E}}(1)\, \hat{\underline{E}}(2) = \hat{Z}_{12} : \hat{\underline{J}}(1)\, \hat{\underline{J}}(2) \quad .$$

Taking the ensemble average we have

$$< \hat{\underset{\sim}{E}}(1)\, \hat{\underset{\sim}{E}}(2) > = Z_{12} : < \hat{\underset{\sim}{J}}(1)\, \hat{\underset{\sim}{J}}(2) > + < Z_{12} : \tilde{T}^{(2)}_{ce} : Z_{12} : \hat{\underset{\sim}{J}}(1)\, \hat{\underset{\sim}{J}}(2) > \quad .$$

The differential equation characterization of $< \hat{\underset{\sim}{E}}(1)\, \hat{\underset{\sim}{E}}(2) >$ is

$$(Y_{12} - V^{(2)}_{ce}) : < \hat{\underset{\sim}{E}}(1)\, \hat{\underset{\sim}{E}}(2) > = < \hat{\underset{\sim}{J}}(1)\, \hat{\underset{\sim}{J}}(2) > + < \tilde{T}^{(2)}_{ce} : Z_{12} : \hat{\underset{\sim}{J}}(1)\, \hat{\underset{\sim}{J}}(2) > \quad .$$

In terms of the fully interacting single particle Green's function Z, this equation reads

$$(Z_1^{-1} Z_2^{-1} - I^{(2)}_e) : < \hat{\underset{\sim}{E}}(1)\, \hat{\underset{\sim}{E}}(2) > = B_{12} : < \hat{\underset{\sim}{J}}(1)\, \hat{\underset{\sim}{J}}(2) > \quad ,$$

where

$$I^{(2)}_e \equiv V^{(2)}_{ce} - V_{ce}(1)\, Z_2^{-1} - Z_1^{-1}\, V_{ce}(2) - V_{ce}(1)\, V_{ce}(2) - < \tilde{V}_{ce}(1)\, \tilde{V}_{ce}(2) > \quad ,$$

and a new operator B_{12} has been introduced by

$$B_{12} : < \hat{\underset{\sim}{J}}(1)\, \hat{\underset{\sim}{J}}(2) > \equiv < (1 + \tilde{T}^{(2)}_{ce}\, Z_{12}) : \hat{\underset{\sim}{J}}(1)\, \hat{\underset{\sim}{J}}(2) > \quad .$$

The proper determination of the operator B_{12} involves a renormalization of the strength of the coupling between $< \hat{\underset{\sim}{E}}(1)\, \hat{\underset{\sim}{E}}(2) >$ and $< \hat{\underset{\sim}{J}}(1)\, \hat{\underset{\sim}{J}}(2) >$.[2]

B. Kinetic Equations for Interacting Quasi-Particles

When the effect of the interaction operator $I^{(2)}_e$ is small it is useful to define renormalized collective modes or wave types by ($Y \equiv Z^{-1}$)

$$Y \cdot \underset{\sim}{E}_\alpha = 0, \qquad \alpha = 1, 2 \ldots$$

for a weakly inhomogeneous system

$$\text{Det } Y(\underset{\sim}{k}, \omega) = 0 \quad ,$$

where $Y(\underset{\sim}{k}, \omega)$ is the Fourier-Laplace amplitude of the fast $(\underset{\sim}{r}, t)$ dependence of Y and the slow space-time dependence of the transform has been suppressed. The above equation defines complex mode frequencies

$$\bar{\omega}_\alpha(\underset{\sim}{k}) = \omega_\alpha(\underset{\sim}{k}) - i\, \gamma_\alpha(\underset{\sim}{k}) \quad , \qquad \alpha = 1, 2 \ldots$$

Because of the implicit, slow space-time dependence of $Y(\underset{\sim}{k}, \omega)$ the mode dispersion relations, $\omega_\alpha(\underset{\sim}{k})$ and growth/decay rates $\gamma_\alpha(\underset{\sim}{k})$ are implicitly slowly varying with position and time. These modes fall into two classes generally referred to as proper and improper or quasi-modes.[1]

Suppose, for simplicity, that there is no background field i.e. $\underset{\sim}{E} = 0$. Then in transform space

$$Y^*(2) \cdot \langle \tilde{\underline{E}}(1)\, \tilde{\underline{E}}^*(2) \rangle$$

$$= Z(1) \cdot \left[B_{12} \langle \hat{\underline{J}}^a(1)\, \hat{\underline{J}}^{a^*}(2) \rangle + I_e^{(2)} \langle \tilde{\underline{E}}(1)\, \tilde{\underline{E}}^*(2) \rangle \right] \qquad (1)$$

where 1, 2... etc. now denote the transform variables $(\underset{\sim}{k}_1, \omega_1)$, $(\underset{\sim}{k}_2, \omega_2)$ etc. and $\hat{\underline{J}}$ has reduced to

$$\hat{\underline{J}}^a \equiv (-e) \int (d\underset{\sim}{v})\, \underset{\sim}{v}\, \hat{f}^a(\underset{\sim}{v}) \quad .$$

The above result is to be compared with Eq. (55) of Ref. (1). Following the procedure outlined in Ref. (1) we now have

$$Y^*(\underset{\sim}{k} + i\nabla,\ \omega - i\partial_t) \cdot \mathcal{S}_{\underset{\sim}{k}\omega}(\underset{\sim}{r}, t) = Z(\underset{\sim}{k}, \omega) \cdot \mathcal{J}_{\underset{\sim}{k}\omega}(\underset{\sim}{r}, t) \quad ,$$

where

$$\mathcal{S}_{\underset{\sim}{k}\omega}(\underset{\sim}{r}, t) \equiv \int \frac{(d\underset{\sim}{k}'')}{(2\pi)^3} \frac{d\omega''}{2\pi} \langle \tilde{\underline{E}}_{\underset{\sim}{k}\omega}\, \tilde{\underline{E}}^*_{\underset{\sim}{k}'\omega'} \rangle\ e^{i(\underset{\sim}{k}'' \cdot \underset{\sim}{r} - \omega'' t)} \quad ,$$

$\underset{\sim}{k}'' \equiv \underset{\sim}{k} - \underset{\sim}{k}'$, $\omega'' \equiv \omega - \omega'$, the slow space-time dependence has now been made explicit, and $\mathcal{J}_{\underset{\sim}{k}\omega}(\underset{\sim}{r}, t)$ is related to the bracketed term in Eq. (1) in the same way that $\mathcal{S}_{\underset{\sim}{k}\omega}(r, t)$ is related to $\langle \tilde{\underline{E}}_{\underset{\sim}{k}\omega}\, \tilde{\underline{E}}^*_{\underset{\sim}{k}\omega} \rangle$ i.e.

$$\mathcal{J}_{\underset{\sim}{k}\omega}(\underset{\sim}{r}, t) \equiv \int \frac{(d\underset{\sim}{k}'')}{(2\pi)^3} \frac{d\omega''}{2\pi} \left[\quad \right]_{\underset{\sim}{k}\omega,\ \underset{\sim}{k}'\omega'} e^{i(\underset{\sim}{k}'' \cdot \underset{\sim}{r} - \omega'' t)} \quad .$$

Integration over the ω-plane contour discussed in Ref. (1) yields

$$Y^*(\underset{\sim}{k} + i\nabla,\ \bar{\omega}_\alpha - i\partial_t) \cdot \mathcal{S}_\alpha(\underset{\sim}{k}, \underset{\sim}{r}, t) = \left(i \frac{\partial Y(\underset{\sim}{k}, \bar{\omega}_\alpha)}{\partial \bar{\omega}_\alpha} \right)^{-1} \cdot \mathcal{J}_{\underset{\sim}{k}\,\bar{\omega}_\alpha}(\underset{\sim}{r}, t) \quad ,$$

where $\mathcal{S}_\alpha(\underset{\sim}{k}, \underset{\sim}{r}, t)$ is the kinetic phase space distribution function for α type quasi-particles of momentum $\underset{\sim}{k}$ and position $\underset{\sim}{r}$. When the mode frequencies, $\bar{\omega}_\alpha$ are sufficiently close to the real axis we have, again following Ref. (1)

$$(\partial_t + \nabla_{\underset{\sim}{k}}\, \omega_\alpha \cdot \nabla - \nabla\omega_\alpha \cdot \nabla_{\underset{\sim}{k}})\, \mathcal{S}_\alpha = 2\gamma_\alpha \mathcal{S}_\alpha - \frac{\mathcal{J}_{\underset{\sim}{k}\omega_\alpha}}{[\partial B(\underset{\sim}{k}, \omega_\alpha)/\partial\omega_\alpha\ ]^2} \quad ,$$

where, for real $(\underset{\sim}{k}, \omega)$

$$Y(\underset{\sim}{k}, \omega) \equiv G(\underset{\sim}{k}, \omega) - iB(\underset{\sim}{k}, \omega) \quad ,$$

$$\gamma_\alpha \equiv G(\underset{\sim}{k}, \omega) \Big/ \frac{\partial B(\underset{\sim}{k}\omega)}{\partial \omega}\Big|_{\omega=\omega_\alpha} ,$$

and, for simplicity, we have assumed an isotropic situation. The above equations are a set of coupled kinetic equations for the quasiparticle distribution functions δ_α

C. Klimontovich Equation for Quasi Particles

When quasi-particles are treated at a level of approximation that includes interactions, a Klimontovich description is desirable. If a Klimontovitch equation can be derived for a stochastic phase space distribution function for quasi particles, all of the techniques developed for dealing with real particles can be applied to quasi particles. In particular quasi particles will be scattered by real particles as well as other quasi particles, localized bound states will occur and also collective oscillations involving quasi particles. The latter will have a spectrum of discrete modes, some of which define new quasi particle types. Presumably the degree of turbulent excitation of a system determines the degree of excitation of the various levels of quasi particles.

The derivation of a Klimontovitch equation for quasi particles follows from

$$(Z_1^{-1} Z_2^{-1} - I_e^{(2)})\ \hat{\underset{\sim}{E}}(1)\ \hat{\underset{\sim}{E}}(2) = (1 + \tilde{T}_{ce}^{(2)} Z_{12})\ \hat{\underset{\sim}{J}}(1)\hat{\underset{\sim}{J}}(2) \quad .$$

We now find

$$(\partial_t + \nabla_{\underset{\sim}{k}}\omega_\alpha - \nabla - \nabla\omega_\alpha \cdot \nabla_{\underset{\sim}{k}})\ \hat{\delta}_\alpha = 2\gamma_\alpha \hat{\delta}_\alpha - \frac{\hat{\mathcal{J}}_{\underset{\sim}{k}\omega_\alpha}}{[\partial B(\underset{\sim}{k}, \omega_\alpha)/\partial\omega_\alpha]^2} ,$$

where $\hat{\delta}_\alpha(\underset{\sim}{k}, \underset{\sim}{r}, t)$ is the stochastic phase space distribution function for quasi-particles of α type. $\hat{\delta}_\alpha$ is derived from

$$\hat{\delta}_{\underset{\sim}{k}\omega}(\underset{\sim}{r}, t) \equiv \int \frac{(d\underset{\sim}{k}'')}{(2\pi)^3} \frac{d\omega''}{2\pi}\ \tilde{\underset{\sim}{E}}_{\underset{\sim}{k}\omega}\ \tilde{\underset{\sim}{E}}^*_{\underset{\sim}{k}'\omega'}\ e^{i(\underset{\sim}{k}''\cdot\underset{\sim}{r} - \omega''t)} ,$$

in the same way that δ_α is derived from $\delta_{\underset{\sim}{k}\omega}$. Notice that δ_α is the ensemble average of $\hat{\delta}_\alpha$. The left hand side of the above equation describes the streaming of α type quasi-particles under the averaged influence of the rest of the system i.e, under the influence of the ensemble averaged "forces". The γ_α term describes the production or absorption of α type quasi particles. The $\hat{\mathcal{J}}$ team is defined by

$$\hat{\mathcal{J}}_{\underset{\sim}{k}\omega}(\underset{\sim}{r}, t) \equiv \int \frac{(d\underset{\sim}{k}'')}{(2\pi)^3} \frac{d\omega''}{2\pi} \Big[\quad \Big]_{\underset{\sim}{k}\omega,\, \underset{\sim}{k}'\omega'} e^{i(\underset{\sim}{k}''\cdot \underset{\sim}{r} - \omega'' t)} ,$$

where (assuming $\underset{\sim}{E} = 0$)

$$\Big[\quad \Big]_{12} = (1 + \tilde{T}^{(2)}_{ce} Z_{12})\, \hat{\underset{\sim}{J}}^{a}(1)\, \hat{\underset{\sim}{J}}^{a*}(2) + I^{(2)}_{e}\, \tilde{\underset{\sim}{E}}(1)\, \tilde{\underset{\sim}{E}}^{*}(2) .$$

Clearly the $\hat{\underset{\sim}{J}}^{a} \hat{\underset{\sim}{J}}^{a}$ term describes the stochastic excitation of the system by the external particle sources $\hat{\eta}$. The $\tilde{\underset{\sim}{E}}\tilde{\underset{\sim}{E}}$ term can be expressed in terms of the Fourier-Laplace transform of $\hat{\delta}_{\underset{\sim}{k}\omega}(\underset{\sim}{r}, t)$ which in turn is a linear combination of all the $\hat{\delta}_{\alpha}(\underset{\sim}{k}, \underset{\sim}{r}, t)$. Thus the above equations are a coupled set for all of the <u>stochastic</u> distribution functions $\hat{\delta}_{\alpha}$ ($\alpha = 1, 2 \ldots$). These equations have a linear appearance. However, there are implicit nonlinearities in $T^{(2)}_{ce}$, Z_{12}, $I^{(2)}_{e}$ etc.

National Science Foundation
ENG76-21829

S. Barone and N. Marcuvitz

Office of Naval Research
N00014-76-C-0176

REFERENCES

1. N. Marcuvitz, J. Math. Phys., Vol. 15, p. 870 (1974).

2. Significant progress has recently been made in understanding the nature of charge renormalization for classical statistical systems. See P. C. Martin, E.D. Siggia and H. A. Rose, Phys. Rev. A, Vol. 8, p. 423 (1973); K. Kawasaki, Prog. Theor. Phys. Vol. 52, p. 1527 (1974); R. Phythian, J. Phys. A: Math. Gen. Vol. 8, p. 1423 (1975); U. Deker and F. Haake, Phys. Rev. A, Vol. 11, p. 2043 (1975) and Phys. Rev. A, Vol. 12, p. 1629 (1975); C. P. Enz and L. Tarrido, Phys. Rev. A, Vol. 14, p. 1258 (1976); R. Phythian, J. Phys. A: Math. Gen. Vol. 10, p. 777 (1977).

HIGH POWER MICROWAVE PROPAGATION THROUGH THE ATMOSPHERE

S. Barone, N. Marcuvitz, R. Pascone and N. Solimene

This is a brief report on the continuing feasibility study to ascertain conditions under which a high power microwave beam propagating through and ionizing the atmosphere can be sufficiently channeled to prevent spreading of the beam energy. Our tentative summary view, based on extrapolation of available breakdown data, is that it is possible to propagate high power nanosecond microwave pulses through the atmosphere provided that the intensity and pulse length are properly selected. Further study is under way to determine whether the self-induced plasma entrained in such pulses can be structured to provide focussing sufficient to offset diffractive spreading of the pulses.

A. CW Propagation

The possibility of self-focussing and self-trapping of radiation fields with beam diameters comparable to or larger than a free space wavelength in fully ionized, asymptotically uniform plasma is well-known. The possibility of a radiation field ionizing an initially neutral atmosphere in such a way as to self-focus or channel is limited by energy considerations, relaxation/growth rates, medium hydrodynamics, etc. Several important conclusions can be drawn on the basis of energy considerations alone. In particular approximately 60 joules are required to singly ionize every molecule in one cubic centimeter of air at sea level. This is of the same order of magnitude as the energy per pulse available from state of the art high power microwave sources. Thus for kilometer path lengths (10^5 cm) say, and beam diameters of the order of a microwave wavelength (~1 cm), the maximum achievable ionization is about one part in 10^5. The density of neutral molecules (~10^{19}/cm^3) is essentially unchanged. For this density of neutrals the electron-neutral collision frequency (ν) is about 10^{12}/sec which exceeds microwave frequencies (ω) by an order of magnitude or more. In this situation the effective linear, non-relativistic dielectric constant is

$$\varepsilon \cong 1 - \left(\frac{\omega_p}{\nu}\right)^2 + i\frac{\nu}{\omega}\left(\frac{\omega_p}{\nu}\right)^2$$

where $\omega_p \cong 10^4\sqrt{N}$ is the plasma frequency in Hz when N is the electron density in particles/cm^3. For the maximum possible fractional ionization of one part in 10^5 or 10^{14} electrons/cm^3, the reduction of the real part of the dielectric constant due to ionization is about 1%. However, for this electron density, the imaginary part of ε is about 0.1 so that the attenuation length due to atmospheric heating is only about 10 wavelengths. In order for the attenuation distance to be comparable to a path length of 10^5 wavelengths the electron density can be no greater than 10^{10}/cm^3. For this situation the real part of the dielectric constant is decreased by 10^{-6}. This is insufficient to cause appreciable self-focussing on a cw basis.

At higher altitudes, say 70,000 feet, atmospheric density is smaller by about a factor of 100, and 60 joules will singly ionize 100 cubic centimeters of air. For kilometer path lengths and centimeter diameter beams, the maximum achievable fractional ionization is 10^{-3}. The electron-neutral collision frequency is 10^{10}/sec which is less than a typical operating frequency, $\omega = 10^{11}$/sec by an order of magnitude. Thus

$$\varepsilon \cong 1 - (\frac{\omega_p}{\omega})^2 + i \frac{\nu}{\omega} (\frac{\omega_p}{\omega})^2$$

and for an attenuation distance of 10^5 wavelengths, the electron density again cannot exceed $10^{10}/cm^3$. For this situation the real part of ε is decreased by 10^{-4}. Again this is insufficient to cause appreciable self-focussing on a cw basis.

These considerations indicate that until very much higher power levels become available, cw channeling is limited to: (1) small beam diameters ($<10^{-2}$ cm) within which full ionization can be achieved or, (2) short pulses, so that the above cw considerations do not apply. Self trapped beams with diameters much less than a free space microwave wavelength are possible and we have studied their properties. The excitation of such channels poses a difficult technological problem.

B. Propagation of Short Pulses

In the atmosphere, electromagnetic waves of power density of the order of a megawatt/cm^2, depending on altitude, produce ionization phenomena that radically modify wave propagation. Ionization gives rise to a space-time dependent plasma that for finite length pulses attenuates the tail of the pulse but does not affect the leading edge because of the finite time for the plasma to build up. Accordingly, for sufficiently short pulses, it is possible to propagate a high power pulse with relatively small attenuation. Since diffractive spreading of such pulses on propagation beyond the Fraunhofer distance results in a decrease in wave intensity, it is of interest to explore whether nonlinear effects in the self-induced plasma can result in sufficient focussing to compensate for wave spreading.

Our preliminary analysis of pulse propagation has been concerned with the determination of optimum pulse size as a function of power density, altitude (atmospheric pressure), etc. We consider at first only ionization, attachment, recombination, and wave attenuation effects for the case of a one-dimensional plane pulse, i.e., we defer consideration of pulse spreading and focussing effects. The model used for our preliminary investigation of the propagation of microwave power through the atmosphere is given in terms of the power density and the electron number density; the transverse spatial dependence has been neglected in this initial work. The normalized equations are:

$$\frac{\partial P}{\partial t} + \frac{\partial P}{\partial z} + \beta P = 0$$

$$\frac{\partial N}{\partial t} = \alpha N - \gamma N^2$$

where P is the power density, relative to the cw breakdown power density and N is the electron density. The unit of distance is such that the velocity of light is unity. The coefficient α is the net ionization frequency:

$$\alpha = \nu_i - \nu_a$$

where ν_i is the ionization frequency and ν_a is the attachment frequency. The coefficient γ is the recombination coefficient. These are empirical functions of E and p where E is the effective field strength and p is the pressure. The attenuation coefficient β is given in terms of the momentum transfer collision frequency, ν, and the electron density, i.e.,

$$\beta = \frac{\omega^2 \nu}{\omega^2 + \nu^2} \frac{N}{N_o} \cong \frac{\omega^2}{\nu} \frac{N}{N_o}$$

where N_o is the electron density corresponding to a plasma frequency equal to ω, i.e.,

$$N_o = \frac{m \omega^2 \epsilon_o}{e^2}$$

In general, the momentum transfer collision frequency is also a function of E and p. However for simplicity our preliminary work has taken ν/p to be constant and equal to 5×10^9 (sec-Torr)$^{-1}$.

The data on ionization and attachment frequencies are often given in terms of the coefficients η_i and η_a respectively where

$$\nu_i = v_d \eta_i$$

$$\nu_a = v_d \eta_a$$

and v_d is the electron drift velocity. The ionization coefficient is well represented by

$$\frac{\eta_i}{p} = A e^{-Bp/E}$$

where for air A = 15(cm-Torr)$^{-1}$ and B = 365 V/(cm-Torr) for E/p in the range 100-800 V/(cm-Torr) (see Reference 1). The data of Harrison and Geballe[2] for E/p in the range 25-65 V/cm-Torr may be fitted with A = 3.12 (cm-Torr)$^{-1}$ and B = 201 V/(cm-Torr). The latter has been used in the calculations summarized in the accompanying

computer plots. Other representations for ν_i have also been used. For example, Mayhan, et al.,[3] use

$$\frac{\nu_i}{p} = f(T,p)\left(\frac{E}{p}\right)^{5.34}$$

where f(T, p) is a given function of temperature and pressure. At room temperature $f = 8.35 \times 10^{-4}$.

The attachment coefficient data is somewhat less certain. The data of Harrison and Geballe[2] for E/p in the range 25-60 V/(cm-Torr) and that of Chatterton and Craggs[4] in the range 2-30 V/(cm-Torr) are in marked disagreement in the overlapping region $25 < E/P < 30$. For the purposes of the present calculation a rough compromise was struck by using

$$\frac{\eta_a}{p} = a + b\left(\frac{E}{p}\right)^2$$

with $a = 5.68 \times 10^{-4}$ (sec-Torr)$^{-1}$ and $b = 3.05 \times 10^{-6}$ (cm-Torr/V)^2/(sec-Torr).

The drift velocity data used is that of Emeleus, Lunt and Meek as given on page 719 of Loeb's book[5] for E/p in the range 20-595 V/(cm-Torr). This data is well summarized by:

$$v_d = 0.125 \times 10^7 \left(\frac{E}{p}\right)^{2/3} \text{ cm/sec}$$

when E/p is given in V/(cm-Torr). Since $v_d = eE/m\nu$, we may infer the momentum transfer collision frequency

$$\frac{\nu}{p} = 1.41 \times 10^9 \left(\frac{E}{p}\right)^{1/3} \text{ (sec-Torr)}^{-1}$$

A cruder fit to the drift velocity data is:

$$v_d = a_o + a_1\left(\frac{E}{p}\right)$$

where $a_o = 1.44 \times 10^7$ cm/sec and $a_1 = 1.15 \times 10^5$ cm^2-Torr/(sec-V). The latter was the form actually used in the calculations performed to date.

The recombination coefficient was taken from the data of Sayers as given in S. C. Brown's book.[6] The value of γ is approximately 2.3×10^{-6} (cm^3/sec) at one atmosphere and decreases to zero as the pressure is lowered. We therefore approximated:

$$\frac{\gamma}{p} = \frac{2.3 \times 10^{-6}}{760} \text{ cm}^3/\text{(sec-Torr)}$$

A summary view of the current state of our analysis is contained in the following series of computer generated graphs displaying the space-time dependent variation of power density in a propagating planar pulse for various parameter ranges. The displays are three-dimensional with power density, POW(z,t), relative to the cw breakdown power density as the ordinate, distance z along the propagation direction as the abscissa (Zf = 0.3 meters), and the third dimension is the time t measured in nanoseconds (Note: one nanosecond corresponds to wave travel of approximately 0.3 meter). The equations as well as the ionization, attachment, etc., parameters used to describe approximately the propagation of a pulsed wave train are indicated in the legend at the top of each graph. The captions are limited to upper case Roman characters and the names of the various parameters are limited to four characters. For these reasons, the following dictionary to the computer graphs is provided:

First line:

D/DT → $\frac{\partial}{\partial t}$

D/DT → $\frac{\partial}{\partial z}$

BETA → β

ALPH → α

GAMA → γ

Note also misprint: N(X, T) should be N(Z, T).

Second line:

BETA=W2/V*(N/NO) means $\beta = \frac{\omega^2}{\nu} \frac{N}{N_o}$

ALPH=NUI-NUA means $\alpha = \nu_i - \nu_a$

NUI=VD*P*a*EXP(-B/(E/P)) means $\nu_i = v_d p A e^{-Bp/E}$

Third line:

NUA=VD*P*(KI+KJ*(E/P)^ 2) means $\nu_a = v_d p[a + b(\frac{E}{p})^2]$

where

KI → a

KJ → b

VD=KU+KV*(E/P) means $v_d = a_o + a_1(\frac{E}{p})$

where

KU → a_o

KV → a_1

GAMA=KK*P means $\gamma = \gamma_1 P$

where

KK → γ

Fourth line:

POW → P_o

OMGA → ω

PRES → p

Fifth line:

N(O) → initial electron density

NO → N_o

NMAX → maximum electron density

In the remaining lines:

WZ/V → $\frac{\omega^2}{\nu}$

EB/P → E_b/p where E_b is effective cw breakdown field strength

ENER → integral of final power density

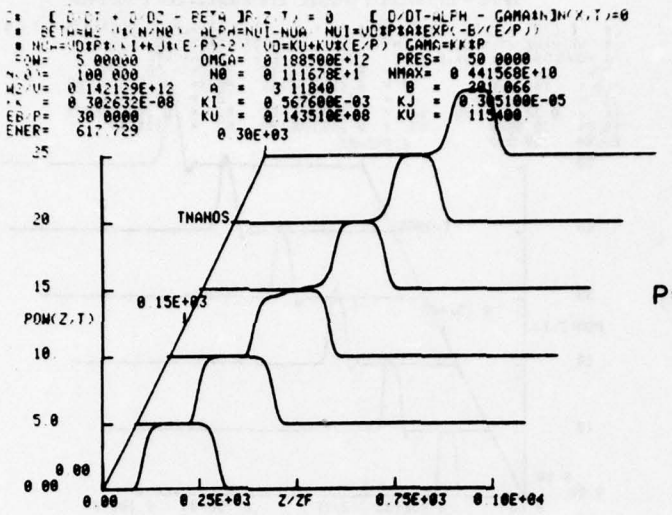
The different sequences display how long pulses are attenuated into short pulses via electron-neutral collisional mechanisms for different initial peak powers, pulse lengths, atmospheric pressures. It should be emphasized that these are preliminary and approximate results; further analysis is continuing.

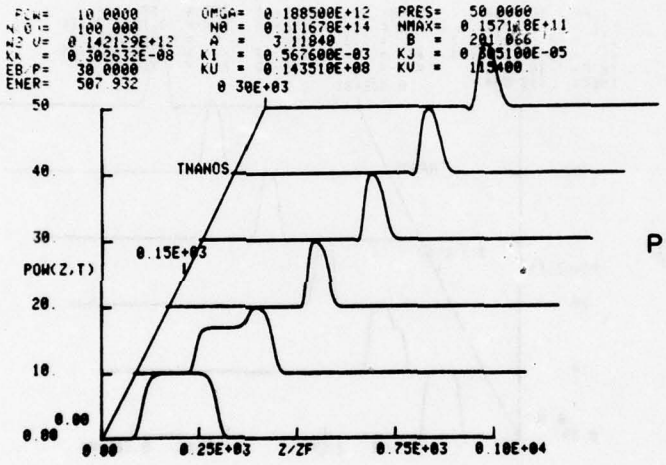
Office of Naval Research
N00014-76-C-0176

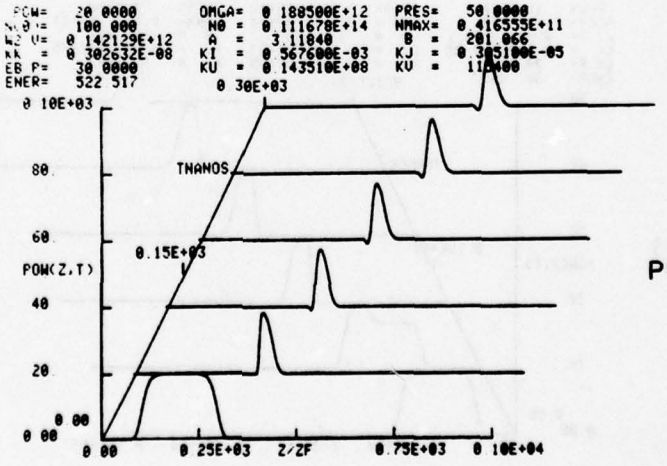
S. Barone, N. Marcuvitz and N. Solimene

REFERENCES

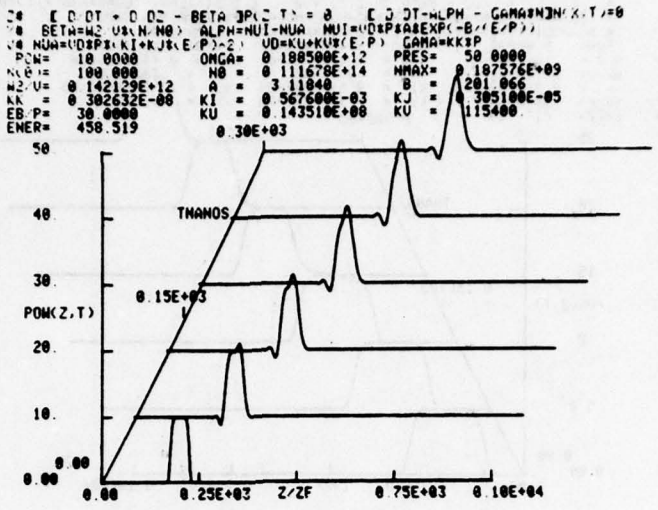
1. A. von Engel, "Encyclopedia of Physics," Vol. 21, Springer (1956).
2. M.A. Harrison and R. Geballe, Phys. Rev., Vol. 91, p. 1 (1953).
3. J.T. Mayhan, et al., J. Appl. Phys., Vol. 42, p. 5362 (1971).
4. P.A. Chatterton and J.D. Craggs, Proc. Phys. Soc., Vol. 85, p. 355 (1965).
5. L.B. Loeb, "Basic Processes of Gaseous Electronics," Univ. of California Press (1961).
6. S.C. Brown, "Basic Data of Plasma Physics," 2nd ed., MIT Press (1966).



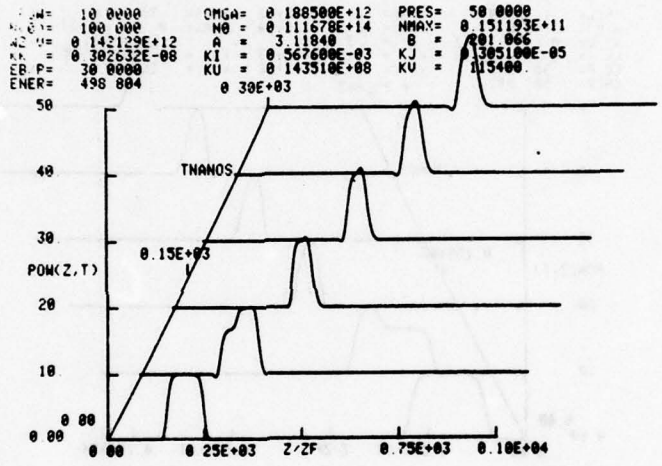




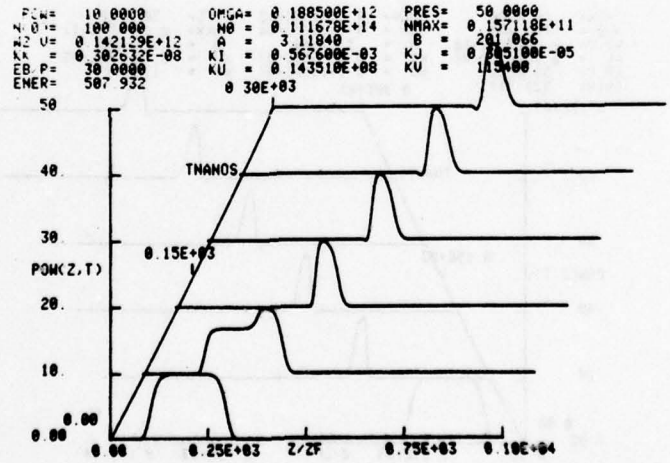
Pulse propagation through atmosphere. Series showing dependence of power.



(a) $\Delta t \approx 50$ ns



(b) $\Delta t \approx 100$ ns



(c) $\Delta t \approx 200$ ns

Pulse propagation through atmosphere. Series showing dependence on pulse width.

A PHENOMENOLOGICAL DESCRIPTION OF INTENSITY DEPENDENT REFLECTION

N. Solimene and M.C. Newstein

The Drude theory of the optical properties of metals is based on a free electron model. Insofar as inter-band transitions, ion-core polarization effects, band structure dependence, electron-electron correlations, non-spherical Fermi-surfaces, etc., may be ignored, this theory yields the complex dielectric constant as a function of the frequency ω

$$\varepsilon(\omega) = 1 - \frac{\omega_p^2}{\omega(\omega + i\nu)}$$

where

$$\omega_p^2 = \frac{Ne^2}{\varepsilon_o m}$$

e = electron change

m = electron mass

N = electron number density

ε_o = permittivity of free space

ν = electron collision frequency

The optical properties are then given by the complex index of refraction $n + ik$ where

$$n = \left\{ \frac{\sqrt{\varepsilon_1^2 + \varepsilon_2^2} + \varepsilon_1}{2} \right\}^{1/2}$$

$$k = \left\{ \frac{\sqrt{\varepsilon_1^2 + \varepsilon_2^2} - \varepsilon_1}{2} \right\}^{1/2}$$

with $\varepsilon = \varepsilon_1 + i\varepsilon_2$. In particular the Fresnel reflection coefficient is given by

$$R = \frac{(n-1)^2 + k^2}{(n+1)^2 + k^2}$$

Thus the metal parameters which determine the optical properties are the effective electron number density, the effective electron mass and the electron collision frequency. For many metals, reasonable choices for these parameters may be made and agreement with experiment obtained for infrared and visible frequencies.

Within the framework of the Drude theory the temperature dependence of the optical properties of pure simple metals is primarily via the temperature dependence of the

collision frequency. The collisions in questions are those between electrons and phonons. The collision frequency increases approximately linearly with temperature between room temperature and the melting point. Thus the collision frequency is expected to increase by a factor of roughly five. Typical values at room temperature are

$$\frac{\omega}{\omega_p} \approx 0.1$$

$$\frac{\nu}{\omega_p} \approx 0.01$$

and R is then higher than 0.9 between room temperature and melting. This is in rough agreement with experiment at low light intensities. However, measurements using laser radiation at intensities of $\sim 10^8$ W/cm^2 indicate reflectivity decreases to values as low as 0.5. The temperature dependent Fresnel reflection coefficient can not account for this and is moreover not applicable since only a thin surface layer is heated.

An attempt to account phenomenologically for the laser reflection coefficients may be based on an intensity dependent dielectric constant. Within the framework of the Drude theory one may introduce intensity dependence by modifying the plasma frequency so that

$$\omega_p^2 = \frac{\omega_{po}^2}{1 + \alpha I}$$

where I is the intensity. Thus the metal becomes less overdense and the penetration of electromagnetic radiation more effective. Qualitative mechanisms for such an effect might involve a metal to non-metal transition mediated by drastic modifications of the electronic band structure in the presence of intense radiation (e.g., high frequency Stark effects). This however will not greatly increase the absorption of radiation, but merely distribute the reflection throughout the affected layer. A second possibility is to modify the collision frequency so that

$$\nu = \nu_o(1 + \beta I)$$

Mechanisms for such an effect might include excitations of plasmons via multi-photon processes and ion-core excitations with results similar to impurity center scattering. Without speculating at this time on the justification of the assumed intensity dependence, it is interesting to evaluate the consequences of these assumptions.

The model used assumes plane wave electromagnetic radiation incident on a constant density metal extending from $z = 0$ to $z = +\infty$. In normalized form, the model equations are

$$\frac{\partial H_x}{\partial t} = \frac{\partial E_y}{\partial z}$$

$$\frac{\partial E_y}{\partial t} = \frac{\partial H_x}{\partial z} + v_y$$

$$\frac{\partial v_y}{\partial t} = -E_y - \nu v_y$$

Inside the metal we may assume

$$E_y = \mathrm{Re}\, \mathcal{E}\, e^{-i\omega t}$$

$$H_x = \mathrm{Re}\, \mathcal{H} e^{-i\omega t}$$

$$v_y = \mathrm{Re}\, v e^{-i\omega t}$$

where the envelope functions $\mathcal{E}$, $\mathcal{H}$ and v are functions of z and very nearly independent of t. We then have

$$\frac{d\mathcal{E}}{dz} = -i\omega \mathcal{H}$$

$$\frac{d\mathcal{H}}{dz} = -i\omega \mathcal{E} - v$$

$$v = \frac{-i\mathcal{E}}{\omega + i\nu}$$

Therefore eliminating v, the second of the above equations becomes

$$\frac{d\mathcal{H}}{dz} = -i\omega\varepsilon\mathcal{E}$$

where

$$\varepsilon = 1 - \frac{1}{\omega(\omega + i\nu)}$$

Note that here ω and ν are normalized and $\omega_p = 1$. When the admittance

$$Y = -\frac{\mathcal{H}}{\mathcal{E}}$$

is introduced, we have

$$\frac{dY}{dz} = -i\omega(Y^2 - \varepsilon)$$

We note that

$$Y(\infty) = \sqrt{\varepsilon}$$

and solve for $Y(z)$ in the interval $0 < z < \infty$. The amplitude reflection coefficient is then given by

$$\rho = \frac{1 - Y(o)}{1 + Y(o)}$$

and the intensity reflection coefficient is

$$R = |\rho|^2$$

The electric field amplitude inside the metal is given by

$$\mathcal{E}(z) = \mathcal{E}(o+)\exp(+ i\omega \int_o^z Y(z')\,dz')$$

where

$$\mathcal{E}(o+) = (1 + \rho)\mathcal{E}_{inc}$$

and $\mathcal{E}_{inc}$ is the incident field amplitude which is prescribed.

Typical results are summarized in Figs. 1 and 2 for $\beta = 1,000$ and $5,000$ respectively. Figure 1(a) shows the incident $|\mathcal{E}_{inc}|^2$ and the transmitted intensity $|\mathcal{E}(o+)|^2$ as well as the reflection coefficient as functions of time. Figure 1(b) shows the intensity $|\mathcal{E}(z)|^2$ for the peak value of $\mathcal{E}_{inc}$. It is to be noted that, for this value of β, the reflection coefficient is appreciably changed, but the field intensity inside the metal is monotonically decreasing. In Fig. 2 the corresponding quantities are shown for $\beta = 5,000$. Notice that the reflection coefficient is drastically reduced and the intensity inside the metal shows a standing wave pattern and indeed $|\mathcal{E}(o+)|^2$ is greater than $|\mathcal{E}_{inc}|^2$. A layer of the metal has in effect become transparent.

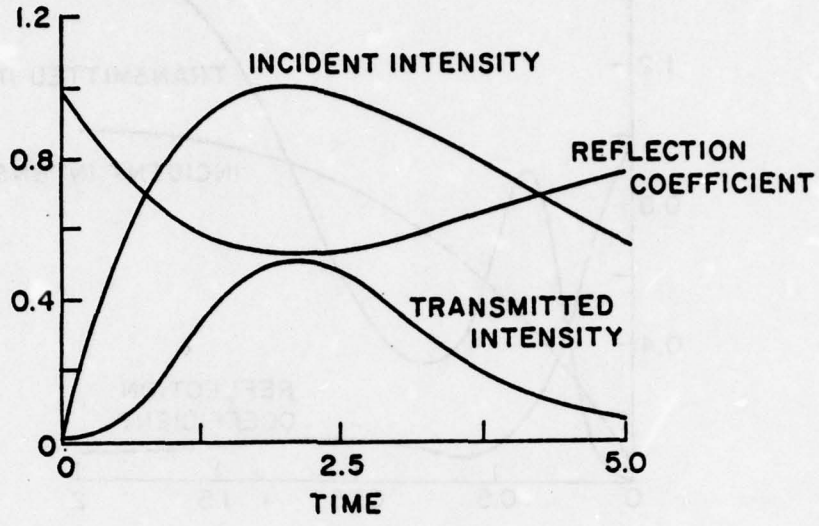
Further work on this model to include thermal effects and a more careful modelling of the metal may be carried out.

National Science Foundation
ENG76-21829

N. Solimene and M.C. Newstein

$\omega = 0.1$ $\nu = 0.01$ KB = 2

$\alpha = 0$ $\beta = 1000$

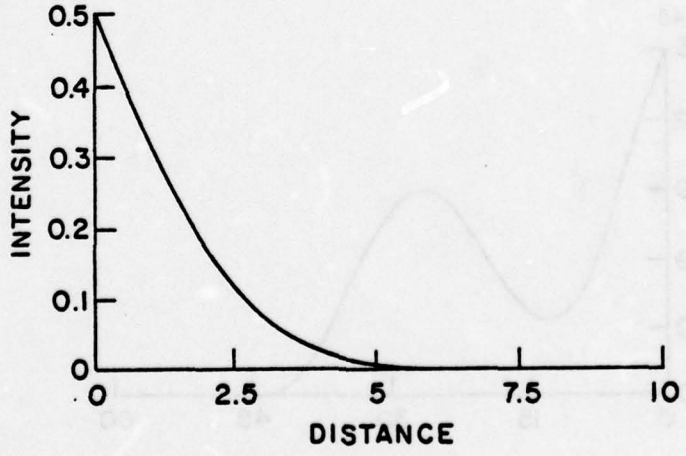


(a)

$\omega = 0.1$ $\nu = 0.01$ KB = 2

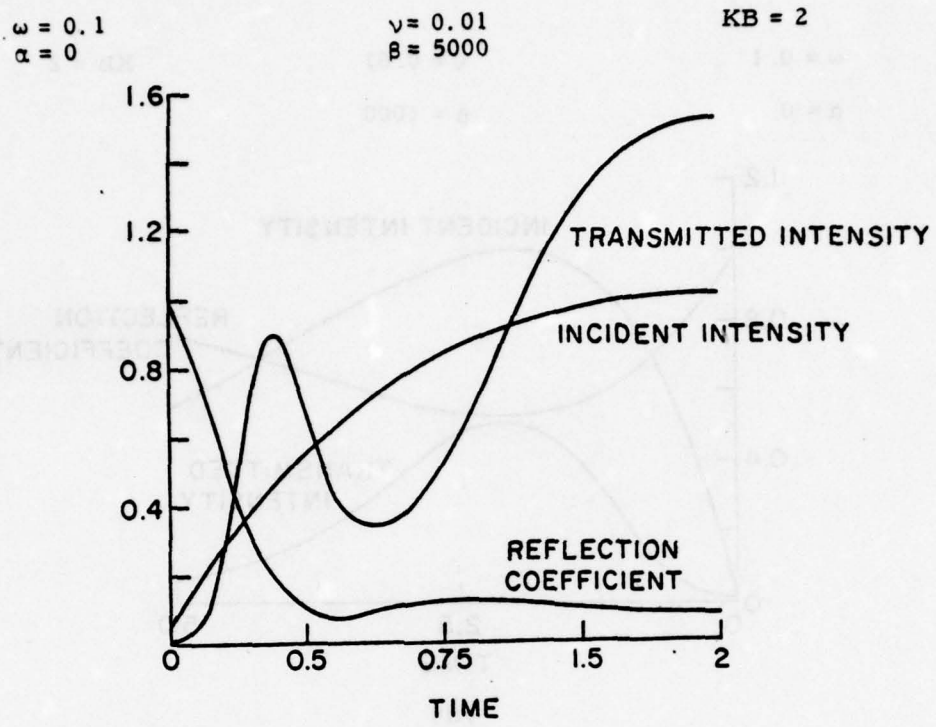
$\alpha = 0$ $\beta = 1000$

R = 0.538347



(b)

Fig. 1. (a) Temporal dependence and (b) Interior spatial dependence for moderate coupling case.



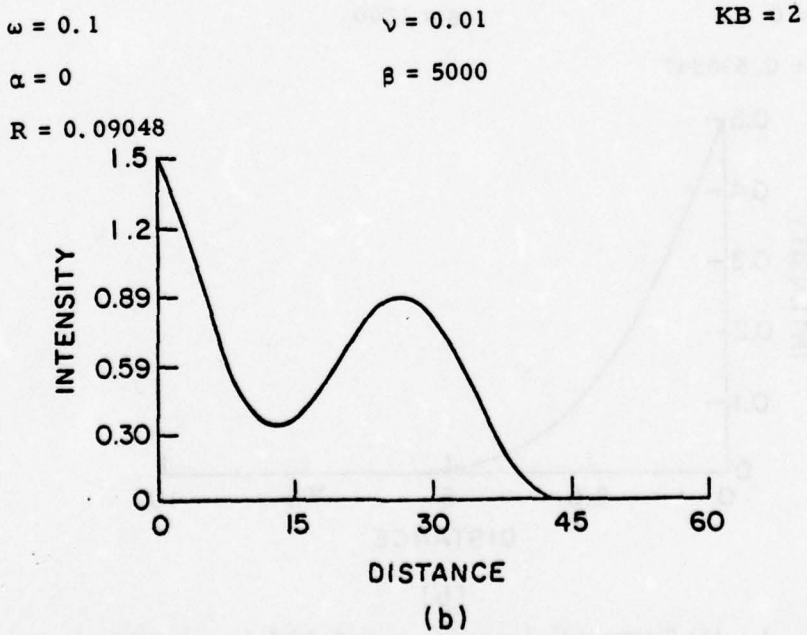


Fig. 2. (a) Temporal dependence and (b) Interior spatial dependence for strong coupling case.

A THEORETICAL STUDY OF INJECTION TUNING OF OPTICAL PARAMETRIC OSCILLATORS

E. S. Cassedy and M. Jain

Injection tuning of pulsed optical parametric oscillators (OPO) has been demonstrated experimentally by Bjorkholm and Danielmeyer[1] in 1969. This technique makes it possible to control the frequency of a high-power, pulsed OPO by injection of low-powered radiation from a frequency controlled source (e.g., an LED). Since thereby the frequency of the OPO can be finely controlled (< 1 cm^{-1} in the IR range), the technique is especially promising for photochemistry applications where power and frequency control are important.

In order to design pulsed OPOs which are injection tuned for single-mode, adjustable-frequency output, the time-dynamic behavior of the entire coupled, multimode OPO system must be understood. Bjorkholm and Danielmeyer[1] gave a good qualitative description of the time-dynamic behavior of successful injection operation. For design criteria, however, a more quantitiative understanding is needed. In the present work we have conducted a computer study of the equations modelling the multimode parametric oscillator. The equations are of the type formulated by Yariv and Louiselle,[2] originally for assumed single-mode operation of the OPO. Here we have generalized these equations to account for coupling into any of the several resonator modes experiencing parametric gain. These generalized equations are as follows:

$$\frac{da_p}{dt} = -i\omega_p a_p - \frac{\omega}{2Q_p} a_p - \sum_{n=1}^{N} K_n a_{2n} a_{2n-1} + i\lambda_p e^{-i\omega t} \tag{1}$$

$$\frac{da_{2n-1}}{dt} = -i\omega_{2n-1} a_{2n-1} - \frac{\omega_{2n-1}}{2Q_s} a_{2n-1} + K_n a^*_{2n} a_p \tag{2}$$

$$\frac{da_{2n}}{dt} = -i\omega_{2n} a_{2n} - \frac{\omega_{2n}}{2Q_i} a_{2n} + K_n a^*_{2n-1} a_p \tag{3}$$

where the (complex) functions are:

$a_p = a_p(t)$ = pump mode amplitude

$a_{2n-1} = a_{2n-1}(t)$ = signal mode amplitude of the nth mode

$a_{2n} = a_{2n}(t)$ = idler mode amplitude for the nth mode

with

ω_{2n-1} = resonant frequency of the nth signal mode

ω_{2n} = frequency of the idler mode,

satisfying: $\omega_p = \omega_{2n-1} + \omega_{2n}$

Q_s = cavity Q factor of signal modes

Q_i = cavity Q factor of idler modes

λ_p = pump-source term; proportional to the incident pump field tangent to the cavity boundary[3]

K_n = coupling factor of nth signal mode

$$= \frac{1}{2}\sqrt{\omega_p \omega_{2n} \omega_{2n-1}} \quad \frac{d}{\epsilon^{3/2} V^{1/2}} \quad \frac{\sin^2(\frac{\Delta kL}{2})}{(\frac{\Delta kL}{2})^2}$$

d = nonlinear susceptibility factor

ϵ = dielectric constant (assumed scalar)

V = volume of the nonlinear crystal

Δk = phase match error in wave-number for the nth mode frequency

L = length of the nonlinear crystal

N = total number of modes

In the model it is possible to consider any number of modes from N = 1 (the single-mode assumption of Yariv and Louiselle) up to the value of N corresponding to the full number under the parametric gain curve: $\sin^2(\frac{\Delta kL}{2}) \Big/ (\frac{\Delta kL}{2})^2$, determined by the phase match characteristics of a particular nonlinear crystal (oriented along a particular axis) and the frequency spacing of the cavity mode resonances. The total number of equations required to model N modes is 2N + 1.

In the computer program the coupling constants have been normalized to that (K_{max}) of the mode with maximum gain and the normalized coupling constant of the nth (non-maximum gain) mode is $\eta_n = K_n / K_{max}$. The mode amplitudes were all normalized to the threshold value:

$$A_{p_{th}} = \frac{\gamma_1 \gamma_2}{2K_{max}} , \qquad \text{where} \begin{cases} \gamma_1 = \frac{\omega_1}{Q_s} \\ \gamma_2 = \frac{\omega_2}{Q_i} \end{cases} \tag{4}$$

which is the pump mode amplitude ($A_p = |a_p|$) at the threshold of growth for a single-mode case.[2] The time scale was normalized to the decay time of the pump mode:

$$T = \frac{2}{\gamma_p} . \tag{5}$$

Using these normalizations in the single-mode (N=1) case, the final steady-state of the equations (following Yariv and Louiselle) for $\frac{da_p}{dT} = \frac{da_1}{dT} = \frac{da_2}{dT} = 0$ results in the amplitudes:

$$|a_p| = 1 \quad , \qquad |a_1| = \left(\frac{\beta-1}{\alpha_1}\right)^{1/2} \quad , \qquad |a_2| = \left(\frac{\beta-1}{\alpha_2}\right)^{1/2}$$

where

$\beta = \lambda_p / \lambda_{p_{th}}$ = the pump driving term normalized to the threshold driving magnitude:

$$\lambda_{p_{th}} = \frac{\gamma_1 \gamma_2 \lambda_p}{4K} \qquad \text{and} \qquad \alpha_1 = \gamma_1 / \gamma_p \quad , \quad \alpha_2 = \gamma_2 / \gamma_b \quad .$$

The Runge Kutta method of numerical integration of differential equations was used to solve the equations. High accuracy solutions were achieved on a digital computer for normalized time increments of $\Delta T = 0.1$ for all mode-number cases tried, including N = 13 (27 complex equations). On Fig. 1 are shown solutions where the pump driving amplitude is twice the threshold value (i.e. $\beta = 2$) for the simple N = 1 case (no injection). The damping constants correspond to a singly-resonant oscillator (SRO), with $\alpha_1 = 0.1$, $\alpha_2 = 1.0$; that is, only the a_1 mode (to be termed the "signal" henceforth) has a decay constant smaller than the pump. The pump-mode amplitude ($|a_p|$) has been given an initial (T = 0) value of zero and the signal and idler mode amplitudes have been given small initial values to correspond to the noise level.

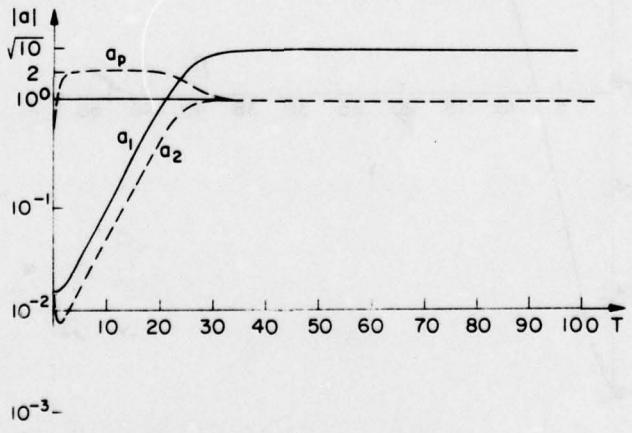


Fig. 1. OPO mode time dynamics (non-injection case)
Parameters: $\alpha_1 = 0.1$, $\alpha_2 = 1.0$, $\beta = 2$

In this figure three regimes are discernible. In the first regime, the pump-mode amplitude (a_p) grows from zero to its undepleted value: $|a_p| = \beta$, and therefore on Fig. 1 it reaches $|a_p| = 2$. Next, noting that on the vertical scale $10^0 = 1$ (i.e., the normalized threshold level), we see that growth of a_1 and a_2 commences as soon as a_p exceeds unity. Once $|a_p| = \beta$, the growth rate of a_1 (and a_2) is just as calculated from the "undepleted-pump" theory, i.e.,

$$e^{+\sqrt{\alpha_1 \alpha_2}\, T} \quad \text{(normalized)} \quad .$$

Finally, the third and final (steady-state) regime is reached as the signal and idler amplitudes become comparable to the pump-mode amplitude and depletion occurs. Note that the final, steady-state, pump amplitude is $|a_p| = 1$, i.e., it settles finally at the threshold level. It can be shown[4] that the final, steady-state solutions are unique for a specified pump source (λ_p, amplitude and phase), being independent of the phases of the initial values of the (small-amplitude) signal (a_1) and idler (a_2) modes.

In order to illustrate injection operation we show on Fig. 2 the results of calculations with one additional mode (a_3) arbitrarily added (the accompanying idler modes

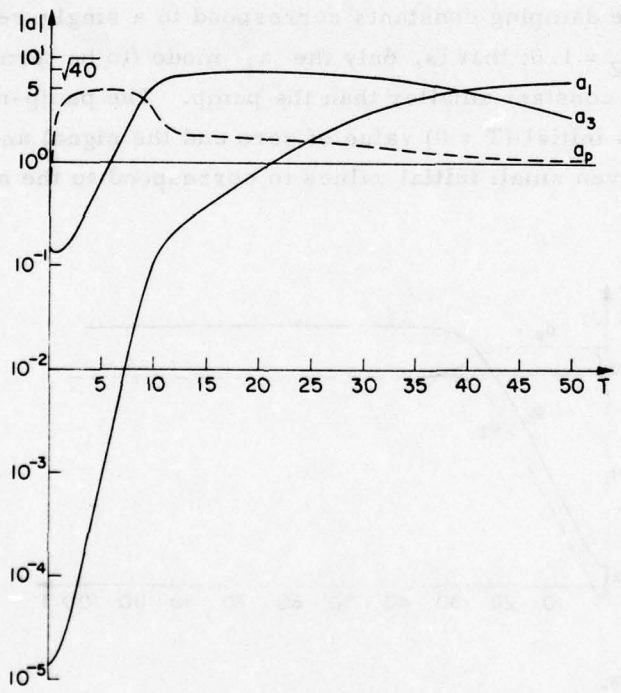


Fig. 2. OPO mode time dynamics (injection case)
Parameters: $\alpha_1 = \alpha_3 = 0.1$, $\alpha_2 = \alpha_4 = 1.0$
$\eta_1 = 1.0$, $\eta_2 = 0.6$
$\beta = 5$

a_2 and a_4 have not been plotted). In Fig. 2 the non-maximum gain mode (a_3) has an initial value of $\sqrt{2} \times 10^{-1}$, due to assumed injection, where as the maximum gain mode (a_1) has an initial value of $\sqrt{2} \times 10^{-5}$ to represent the noise level. We see that a quasi-steady-state is established during the period T ≃ 12-27, during which the pump mode amplitude has been depleted to the (normalized) value ($|a_p| = \frac{1}{\eta_2} = 1.67$) by the injected mode ($a_3$) which dominates in this regime. Subsequently, the fastest growing mode (a_1) attains magnitudes sufficient to further deplete the pump and the final steady-state ($a_p = 1$) is reached essentially by T = 50. The a_3 mode has stated to decay, once the pump amplitude is depleted below its ($|a_p| = 1.67$) threshold value (for $T \gtrsim 27$) and continues to decay thereafter to insignificant magnitudes.

The mode dynamics shown on Fig. 2, represents "successful" injection operation, in the sense that an essentially-pure* frequency of ω_3 could be expected out of such an oscillator over a (normalized) time period of about 15. Such an output would be shifted from the ω_1 frequency which would be emitted in the absence of injection at ω_3, or which is emitted in the final steady-state. Some OPO parameters do not lead to successful injection operation, however, an example of which is shown on Figure 3.

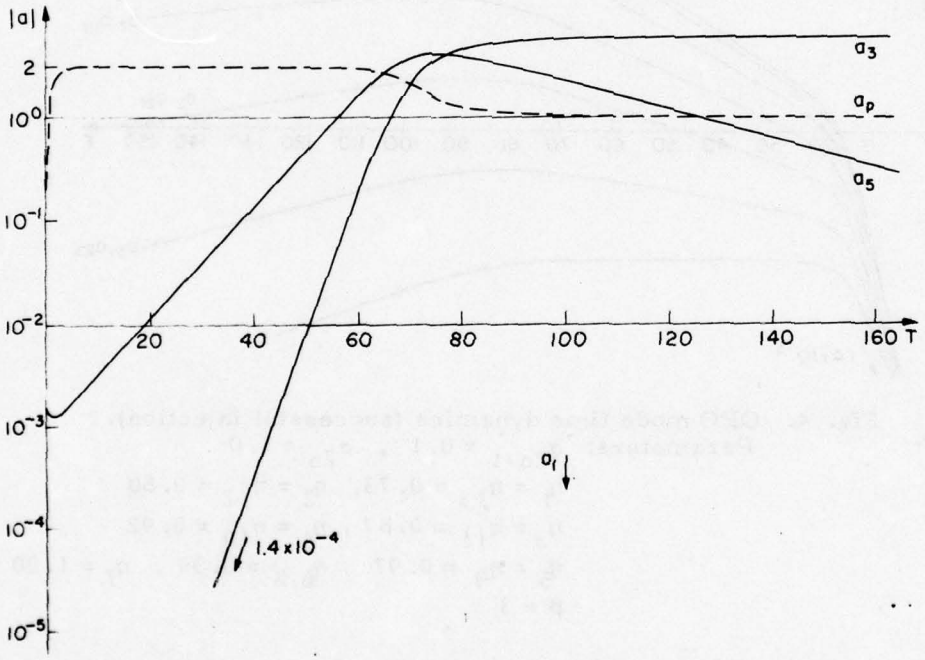


Fig. 3. OPO mode time Dynamics (unsuccessful injection).
Parameters: $\alpha_1 = \alpha_3 = \alpha_5 = 0.055$
$\alpha_2 = \alpha_4 = \alpha_6 = 1.0$
$\eta_1 = \eta_3 = 0.637$, $\eta_2 = 1.0$
$\beta = 2$

* We ignore the idler frequencies, since they are in another wavelength band for typical OPOs.

Here, an N = 3 case, the injection is on mode a_5, which is a non maximum gain mode (along with a_1). The mode labeled a_3 is here the maximum gain mode and may be seen to overtake the injected mode before a quasi-steady-state can be established. As a result the pump becomes depleted to the $|a_p| = 1$ resulting in the decay of mode a_5 and "unsuccessful" injection. Further calculations[4] show that successful injection can be achieved for the oscillator parameters of Fig. 3 if the pump source* is increased. Such a prescription for successful injection was not, however, found effective in any case where injection was attempted on a mode having a relative coupling constant (η_n) less than 0.6.

On Fig. 4 is shown a case of successful injection, where the injected mode (a_{25}) has a relative coupling constant η_{13}=0.73 and the (relative) pump strength is $\beta = 3$.

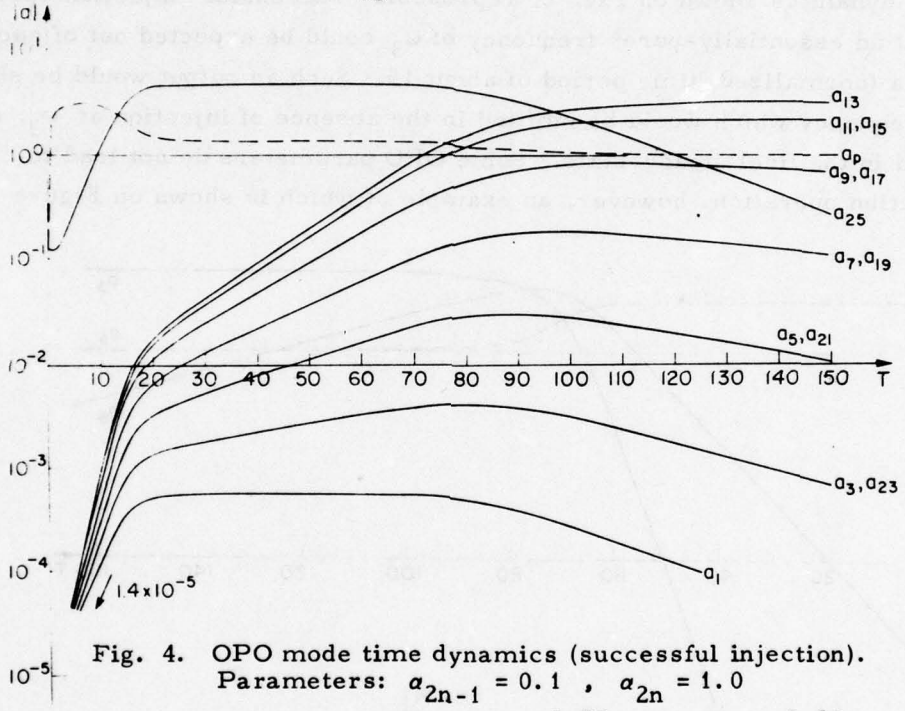


Fig. 4. OPO mode time dynamics (successful injection).
Parameters: $\alpha_{2n-1} = 0.1$, $\alpha_{2n} = 1.0$
$\eta_1 = \eta_{13} = 0.73$, $\eta_2 = \eta_{12} = 0.80$
$\eta_3 = \eta_{11} = 0.87$, $\eta_4 = \eta_{10} = 0.92$
$\eta_5 = \eta_9 = 0.97$, $\eta_{6,8} = 0.99$, $\eta_7 = 1.00$
$\beta = 3$

* The pump source parameter (λ_p) corresponds to the field strength of the (external) pumping laser.

A quasi-steady-state is seen to be established for over 50 time units. This figure also illustrates the (small) influence of a large number of off-peak-gain modes, as would be present in an actual OPO. The curves labelled with two modes (e.g., the a_{11}, a_{15} curve second from the top on the right) represent modes whose frequencies are disposed symmetrically either side of the maximum gain mode, and thus these modes have identical gain characteristics. In any case, these modes, if not injected, contribute only to a minor extent to the depletion dynamics and all decay once the peak-gain mode depletes the pump mode.

Throughout these calculations one parameter stands out as a consistent measure of success for injection tuning. That parameter is $\eta = K_{inject}/K_{max}$, the coupling constant of the injected mode relative to that of the highest-gain mode. Several injection criteria can be set down:

a) if $\eta < 0.6$, then successful injection cannot be expected

b) an increase in pump strength (i.e., $\beta >> 1$) will not result in successful injection, when $\eta \leq 0.6$

c) if $\eta > 0.6$, an increase in pump strength can improve injection

d) for low pump strength ratios (e.g., $\beta = 1.5$), coupling constants considerably higher than 0.6 (e.g., $\eta > 0.85$) are required for successful injection

e) an increase in cavity Q or cavity-reflectance appears to improve injection, when the operation is already at least in the marginal success range.

The η parameter is not a directly measurable quantity for the experimenter. It nonetheless can be estimated from the gain taper once the total bandwidth of the OPO cavity is known, by a method used in this study.[4] The results get translated into the number of modes "off center" which can be used for injection tuning.[4]

Los Alamos Scientific Laboratory
Purchase Order: NP6-21964-1

E. S. Cassedy

REFERENCES

1. J.E. Bjorkholm and H.G. Danielmeyer, "Frequency Control of a Pulsed Optical Parametric Oscillator by Radiation Injection," Appl. Phys. Lett., Vol. 15, pp. 171-173 (1969).

2. A. Yariv and W.H. Louiselle, "Theory of the Optical Parametric Oscillator," IEEE Journ., Vol. QE-2, pp. 418-424 (1966).

3. J.C. Slater, "Microwave Electronics," Chap. IV, Van Nostrand, NY.

4. E.S. Cassedy and M. Jain, "A Theoretical Study of Injection Tuning of Optical Parametric Oscillators," to be published.

PARAMETRIC EXCITATION OF COUPLED PLASMA WAVES

B. R. Cheo, C. Hechtman, S. P. Kuo, T. Q. Yip

A. Introduction

Parametric decay instabilities in plasmas have been studied extensively. In one recent review article by Porkolab,[1] over one hundred references are cited. For non-magneto plasmas, the theoretical efforts are typically represented by that of Nishikawa[2] whose basic approach is to use the hydrodynamic equations to derive the coupled mode equations. For magneto plasma, the method generally used is a transformation by Aliev et al.[3] on the Vlasov equation as shown by the work of Porkolab.[4] In this case, the quasi static approximation $\underline{E} = -\nabla \Phi$ has been used and hence valid for longitudinal waves. In this report we present a general analysis based on the Hamiltonians of the decay waves. The approach is valid for all wave types. In the limiting case of $B_o \to 0$ the result is reduced to that of Nishikawa.[2] In the limit of longitudinal waves the results of Porkolab[1,4] are recovered.

B. Coupled Mode Equations

We consider two normal modes designated by S and L (e.g., a phonon and a plasmon) with frequencies ω_s' and ω_ℓ' in the linear regime, wave numbers $\pm k$ (uniform pump field assumed), and reduced masses for the two species M and m. They can be taken as ion and electron masses if S is dominated by ions and L by electrons. The unperturbed Hamiltonian density of each mode is given by

$$H_{os}^{(k)} = \frac{1}{2} \sum_\sigma (K_\sigma^2(\underline{k})/M + M\,\omega_s'^2\, Q_\sigma^2(\underline{k}))$$

$$H_{o\ell}^{(k)} = \frac{1}{2} \sum_\sigma (\mathcal{K}_\sigma^2(\underline{k})/m + m\,\omega_\ell'^2\, U_\sigma^2(\underline{k})) \qquad \sigma = 1, 2, 3 \text{ Cartesian components} \qquad (1)$$

where $\underline{K}$, $\underline{\mathcal{K}}$; $\underline{Q}$, $\underline{U}$ are respectively the canonical momentum and displacement variables. Assuming a static magnetic field $\underline{B} = B_o \hat{z}$, these variables are normalized to satisfy the Poisson bracket relations:

$$\{Q_\sigma(\underline{k}),\ K_\beta(\underline{k}')\} = \delta(\underline{k}, \underline{k}')\,\delta_{\sigma,\beta} \quad \text{and} \quad \{U_\sigma(\underline{k}),\ \ _\beta(\underline{k}')\} = \delta(\underline{k}, \underline{k}')\,\delta_{\sigma,\beta}$$

and are related to the modal fields as:

$$\underline{E}_s(\underline{k}) = -\frac{M}{e\sqrt{n_o}}\,\omega_s^2(\underline{k})\,\underline{Q}(\underline{k}) - \frac{\Omega_i}{e\sqrt{n_o}}\,\underline{K}(\underline{k}) \times \hat{z}$$

$$\underline{E}_\ell(\underline{k}) = \frac{m}{e\sqrt{n_o}}\,\omega_\ell^2(\underline{k})\,\underline{U}(\underline{k}) + \frac{\Omega_e}{e\sqrt{n_o}}\,\underline{\mathcal{K}}(\underline{k}) \times \hat{z} \qquad (2)$$

where ω_s and ω_ℓ are the actual (shifted) oscillating frequencies of the respective modes, Ω_e and Ω_i are cyclotron frequencies for masses m and M with the appropriate sign of the charges, and n_o is the background density. When the plasma is pumped by an external field $\underline{E}_p = \underline{E}_1 + \underline{E}_2 = 2\underline{E}_{10} \sin\omega_o t + 2\underline{E}_{20} \cos\omega_o t$, coupling of the modes will take place. The Hamiltonians will be perturbed through the modulation of the electrorestrictive polarizability tensor α_{ij}. In the first order Taylor's series expansion, we have (Einstein summation convention is understood):

$$\alpha_{ij} \cong \alpha_{ij}^o + \left(\frac{\partial \alpha_{ij}}{\partial E_{s\sigma}}\right)_o E_{s\sigma} + \left(\frac{\partial \alpha_{ij}}{\partial E_{\ell\alpha}}\right)_o E_{\ell\sigma} \tag{3}$$

Keeping terms of frequencies ω_o, ω_ℓ and ω_s the induced polarization due to the pump is expressed as

$$P_i = \alpha_{ij} E_{pj} = \alpha_{ij}^o E_{pj} + \left(\frac{\partial \alpha_{ij}}{\partial E_{s\sigma}}\right)_o E_{s\sigma} E_{pj} + \left(\frac{\partial \alpha_{ij}}{\partial E_{\ell\sigma}}\right) E_{\ell\sigma} E_{pj} \tag{4}$$

Because of the conservation of frequencies $\omega_o = \omega_\ell + \omega_s$, the second term may be designated by $\delta P_{\ell i}$ and the third by δP_{si}. The interaction Hamiltonians of S and L are thus $H_s'^{(k)} = -\delta P_{si}^{(k)} E_{si}(\underline{k})$ and $H_\ell'^{(k)} = -\delta P_{\ell i}^{(k)} E_{\ell i}(\underline{k})$, with:

$$\begin{aligned} \delta P_{si}^{(k)} &= \left(\frac{\partial \alpha_{ij}}{\partial E_{\ell\sigma}^+(\underline{k})}\right)_o E_{\ell\sigma}^+(\underline{k}) E_{pj}^- + \text{c.c.} = \delta P_{si}^- + \text{c.c.} \\ \delta P_{\ell i}^{(k)} &= \left(\frac{\partial \alpha_{ij}}{\partial E_{s\sigma}^-(\underline{k})}\right) E_{s\sigma}^- E_{pj}^+ + \text{c.c.} = \delta P_{\ell i}^+ + \text{c.c.} \end{aligned} \tag{5}$$

The superscripts "$\pm$' here and elsewhere designate recpectively the $\exp(\pm i\omega t)$ components of the oscillating fields. The total Hamiltonian of each mode is then the sum of the unperturbed and interaction Hamiltonians $H = H_o + H'$. It is known that the equation of motion for any canonical variable A is $\dddot{A} + 2\Gamma\ddot{A} + \Gamma^2\dot{A} = \{\{A, H\}, H\}$, where Γ^{-1} is the phenomenological relaxation time. We can apply this to derive the equations of motion for the four variables. Using Eq. (2), we then obtain the coupled mode equations:

$$\begin{aligned} \ddot{\underline{E}}_s^-(\underline{k}) + 2\Gamma_s \dot{\underline{E}}_s^- + \omega_s^2(\underline{k})\,\underline{E}_s^- &= \frac{M}{n_o e^2}\left[\omega_s^4(\underline{k})\,\delta\underline{P}_s^-(\underline{k}) + \omega_s'^2(\underline{k})\,\Omega_i^2(\hat{z} \times \delta\underline{P}_s^-(\underline{k}))\right] \\ \ddot{\underline{E}}_\ell^+(\underline{k}) + 2\Gamma_\ell \dot{\underline{E}}_\ell^+ + \omega_\ell^2(\underline{k})\,\underline{E}_\ell^+ &= \frac{m}{n_o e^2}\left[\omega_\ell^4(\underline{k})\,\delta\underline{P}_\ell^+(\underline{k}) + \omega_\ell'^2(\underline{k})\,\Omega_e^2(\hat{z} \times \delta\underline{P}_\ell^+(\underline{k})) \times \hat{z}\right] \end{aligned} \tag{6}$$

where $\omega^2_{\ell,s} = \omega'^2_{\ell,s} + \Gamma^2_{\ell,s}$ the shifted oscillating frequencies. It is seen that $\delta \underline{P}$ on the RHS of Eq. (6) provides the coupling between S and L as shown in Equation (5).

C. Coupling Coefficients

To obtain the coupling coefficients $\frac{\partial \alpha_{ij}}{\partial \underline{E}_{s,\ell}}$ we consider the distribution function we consider the distribution function $f_\alpha(\underline{v}_\alpha, \underline{r}_\alpha, t)$ for species α which satisfies the Vlasov equation

$$\partial_t f_\alpha + \underline{v}_\alpha \cdot \partial_{\underline{r}_\alpha} f_\alpha + \frac{e_\alpha}{m_\alpha}\left[\underline{E} + \frac{1}{c}\,\underline{v}_\alpha \times \underline{B}_o\right] \cdot \partial_{\underline{v}_\alpha} f_\alpha = 0 \tag{7}$$

Using a procedure similar to that of Aliev et al.,[4] we first transform Eq. (7) to the oscillating frame of reference: $\underline{v} = \underline{v}_\alpha - \frac{e_\alpha}{m_\alpha}\int_{-\infty}^{t} dt' \underline{\underline{R}}_\alpha(t-t') \cdot \underline{E}_p(t')$ and $\underline{r} = \underline{r}_\alpha - \frac{e_\alpha}{m_\alpha}\int_{-\infty}^{t} dt' \int_{-\infty}^{t'} dt'' \underline{\underline{R}}_\alpha(t'-t'') \cdot \underline{E}_p(t'')$ where $\underline{\underline{R}}_\alpha(t) = (\hat{x}\hat{x} + \hat{y}\hat{y}) \cos \Omega_\alpha t + (\hat{x}\hat{y} - \hat{y}\hat{x}) \sin \Omega_\alpha t + \hat{z}\hat{z}$ ($\Omega_\alpha = e_\alpha B_o / m_\alpha c$) and $\underline{E}_p(t)$ is the pump field. Let $f_\alpha = f^{(o)}_\alpha + \psi_\alpha$ and $\underline{E} = \underline{E}_p + \delta \underline{E}$ with $f^{(\delta)}_\alpha$ being a spatially uniform function, $|\psi_\alpha| << |f^o_\alpha|$ and $|\delta \underline{E}| << |\underline{E}_p|$. Collecting zeroth and first order terms of the transformed Eq. (7), we have

$$\partial_t f^{(o)}_\alpha(\underline{v}, t) + \Omega_\alpha \underline{v} \times \hat{z} \cdot \partial_{\underline{v}} f^{(o)}_\alpha(\underline{v}, t) = 0 \tag{8}$$

and

$$(\partial_t + \underline{v} \cdot \partial_{\underline{r}} + \Omega_\alpha \underline{v} \times \hat{z} \cdot \partial_{\underline{v}})\ \psi_\alpha = -\frac{e_\alpha}{m_\alpha}\,\delta \underline{E} \cdot \partial_{\underline{v}} f^{(o)}_\alpha \tag{9}$$

Assuming that $f^{(o)}_\alpha = \left(\frac{m_\alpha}{2\pi T_\alpha}\right)^{3/2} \exp\left[-(v^2_\perp + v^2_{11})\, m_\alpha / 2T_\alpha\right]$, a Maxwellian in $\underline{v}$, and that the spatial dependence of ψ_α and $\delta \underline{E}$ take the form $\exp(i\underline{k} \cdot \underline{r}_\alpha)$, we can integrate Eq. (9) along an unperturbed trajectory: $\dot{\underline{r}} = \underline{v}$, $\dot{\underline{v}} = \Omega_\alpha \underline{v} \times \hat{z}$. Along such trajectories $\underline{v}(t') = \underline{\underline{R}}(t'-t) \cdot \underline{v}(t)$ and $\underline{r}(t') = \underline{r}(t) + \Omega^{-1}_\alpha \underline{\underline{L}}_\alpha(t'-t) \cdot \underline{v}(t)$. where $\underline{\underline{L}}_\alpha(t) = \Omega_\alpha \int_o^t \underline{\underline{R}}_\alpha(\tau)\, d\tau$. Along this trajectory the LHS of Eq. (9) becomes an exact differential. Integrating the RHS along the trajectory thus yields ψ_α. The total induced current, and hence the polarization, can be then calculated from $\underline{J} = \sum_\alpha \int \underline{v}_\alpha \psi_\alpha d\underline{v}_\alpha$. For simplicity, we choose the wave vector $\underline{k}$ to be in the x-z plane $\underline{k} = k_\perp \hat{x} + k_{11} \hat{z}$, the integration can be carried out through the transformation of the variables in the $(\underline{r}, \underline{v})$ frame, yielding the total dielectric tensor. The linear part can be shown to agree with the well established form.[5] The nonlinear part can be best described by the induced polarization $\delta \underline{P}^-_s$ and $\delta \underline{P}^+_\ell$. After much algebra[6,7] we obtain

$$\delta \underline{P}_s^-(\underline{k},t) = \frac{i}{\omega_s} \sum_\alpha \frac{n_o e_\alpha^3 \omega_o}{m_\alpha T_\alpha(\omega_o^2 - \Omega_\alpha^2)} \underline{E}_\ell^+(\underline{k},t) \cdot \sum_{n=-\infty}^{+\infty} \frac{\Omega_\alpha}{k_\perp(\omega_\ell + n\Omega_\ell)} (1 - W(Z_{\ell n}^\alpha))(n\Lambda_n(\beta_\alpha)\hat{x}$$

$$+ i\beta_\alpha \Lambda_n'(\beta_\alpha)\hat{y} + \frac{W(Z_{\ell n}^\alpha)}{k_\parallel} \Lambda_n(\beta_\alpha)\hat{z}) \; \underline{\underline{A}}_\alpha + \frac{\beta_\alpha \Omega_\alpha^2}{k_\perp^2 \omega_o \omega_\ell} \sum_n \frac{Z_{\ell o}^\alpha}{Z_{\ell n}^\alpha} \underline{\underline{\Pi}}(\beta_\alpha, Z_{\ell n}^\alpha; n)(1 - W(Z_{\ell n}^\alpha))$$

$$-(Z_{\ell o}^\alpha)^2 \hat{z}\hat{z} \; \underline{k} \cdot \underline{\underline{A}}_\alpha + \frac{\beta_\alpha \Omega_\alpha^2}{k_\perp^2 \omega_o \omega_s} \sum_n \frac{Z_{so}^\alpha}{Z_{sn}^\alpha} \underline{\underline{\Pi}}(\beta_\alpha, Z_{sn}^\alpha; n)(1 - W(Z_{sn}^\alpha)) - (Z_{so}^\alpha)^2 \hat{z}\hat{z} \; \underline{k} \cdot \underline{\underline{A}}_\alpha \cdot \underline{E}_p^- \tag{10}$$

$$\delta \underline{P}_\ell^+(\underline{k},t) = -\frac{i}{\omega_\ell} \sum_\alpha \frac{n_o e_\alpha^3 \omega_o}{m_\alpha T_\alpha(\omega_o^2 - \Omega_\alpha^2)} \underline{E}_s^-(\underline{k},t) \cdot \sum_n \frac{\Omega_\alpha}{k_\perp(\omega_s - n\Omega_\alpha)} (1 - W(Z_{sn}^\alpha))(n\Lambda_n(\beta_\alpha)\hat{x}$$

$$+ i\beta_\alpha \Lambda_n'(\beta_\alpha)\hat{y}) - \frac{W(Z_{sn}^\alpha)}{k_\parallel} \Lambda_n(\beta_\alpha)\hat{z} \; \underline{\underline{A}}_\alpha^* - \frac{\beta_\alpha \Omega_\alpha^2}{k_\perp^2 \omega_o \omega_s} \sum_n \frac{Z_{so}^\alpha}{Z_{sn}^\alpha} \underline{\underline{\Pi}}(\beta_\alpha, Z_{sn}^\alpha; n)(1 - W(Z_{sn}^\alpha))$$

$$-(Z_{so}^\alpha)^2 \hat{z}\hat{z} \; \underline{k} \cdot \underline{\underline{A}}_\alpha^* - \frac{\beta_\alpha \Omega_\alpha^2}{k_\perp^2 \omega_o \omega_\ell} \sum_n \frac{Z_{\ell o}^\alpha}{Z_{\ell n}^\alpha} \underline{\underline{\Pi}}(\beta_\alpha, Z_{\ell n}^\alpha; n)(1 - W(Z_{\ell n}^\alpha)) - (Z_{\ell o}^\alpha)^2 \hat{z}\hat{z} \; \underline{k} \cdot \underline{\underline{A}}_\alpha^* \cdot \underline{E}_p^+ \tag{11}$$

where $\beta_\alpha = \dfrac{k_\perp^2 T_\alpha}{m_\alpha \Omega_\alpha^2}$; $Z_{sn}^\alpha = +\dfrac{\omega_s - n\Omega_\alpha}{k_\parallel (T_\alpha/m_\alpha)^{1/2}}$, $Z_{\ell n}^\alpha = \dfrac{\omega_\ell + n\Omega_\alpha}{k_\parallel (T_\alpha/m_\alpha)^{1/2}}$,

$\Lambda_n(\beta) = I_n(\beta) e^{-\beta}$ (I_n is the modified Bessel function), $\underline{\underline{\Pi}}(\beta, Z_n; n) = \dfrac{n^2}{\beta} \Lambda_n(\beta)\, \hat{x}\hat{x}$

$$+ \left[\frac{n^2}{\beta} \Lambda_n - 2\beta \Lambda_n'\right] \hat{y}\hat{y} + Z_n^2 \Lambda_n \hat{z}\hat{z} - in\Lambda_n'(\hat{x}\hat{y} - \hat{y}\hat{x}) + \frac{k_\perp \Omega}{|k_\perp \Omega|} \frac{n}{\sqrt{\beta}} Z_n \Lambda_n(\hat{x}\hat{z} + \hat{z}\hat{x})$$

$$+ \frac{ik_\perp \Omega}{|k_\perp \Omega|} \sqrt{\beta}\, Z_n \Lambda_n'(\hat{y}\hat{z} + \hat{z}\hat{y}), \quad W(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} -\frac{x}{x - Z - io} \exp\left(-\frac{x^2}{2}\right) dx$$

and $\underline{\underline{A}}_\alpha = \hat{x}\hat{x} + \hat{y}\hat{y} + i\dfrac{\Omega_\alpha}{\omega_o}(\hat{x}\hat{y} - \hat{y}\hat{x}) + \dfrac{\omega_o^2 - \Omega_\alpha^2}{\omega_o^2} \hat{z}\hat{z}$. $\underline{\underline{A}}_\alpha^*$ is the complex conjugate of $\underline{\underline{A}}_\alpha$.

Using Eqs. (10), (11) and (6), one can calculate the threshold fields for various decay processes.

D. Comparison with Existing Results

Because only longitudinal waves are considered, we can set $\underline{E}_s^- = i\underline{k}\,\Phi_s^-$ and $\underline{E}_\ell^+ = -i\underline{k}\Phi_\ell^+$, and rewrite Eq. (6) as

$$\ddot{\Phi}_s^- + 2\Gamma_s\,\dot{\Phi}_s^- + \omega_s^2\,\Phi_s^- = \frac{iM}{n_o e^2 k^2}\,\underline{k}\cdot\left[\omega_s^4(\underline{k})\,\delta\underline{P}_s^-(\underline{k}) + \omega_s'^2(\underline{k})\,\Omega_i^2(\hat{z}\times\delta\underline{P}_s^-(\underline{k}))\times\hat{z}\right] \quad (12)$$

$$\ddot{\Phi}_\ell^+ + 2\Gamma_\ell\,\dot{\Phi}_\ell^+ = \frac{im}{n_o e^2 k^2}\,\underline{k}\cdot\omega_\ell^4(\underline{k})\,\delta\underline{P}_\ell^+(\underline{k}) + \omega_\ell'^2(\underline{k})\,\Omega_e^2(z\times\hat{o}\,\underline{P}_\ell^+(\underline{k}))\times\hat{z} \quad (13)$$

We consider two cases when comparison can be made: (1) $B_o = 0$ or the pump wave is the ordinary mode with decay waves: Langmuir and ion acoustic waves; and (2) pump wave is the extraordinary mode, $|\Omega_e| >> \omega_{pe}$, with decay waves: upper hybrid and electrostatic ion cyclotron waves, $(m/M)^{1/2} < k_\parallel/k_\perp << 1$; or upper hybrid and low hybrid waves, $k_\parallel/k_\perp << (m/M)^{1/2}$: (1) Let $\underline{E}_p(t) = 2E_{20}\cos\omega_o t\,\hat{z} = \underline{E}_2$ and $\underline{k} = k\hat{z}$. Substitution of these in Eqs. (10) to (13) yields

$$\ddot{\Phi}_s^- + 2\Gamma_s\,\dot{\Phi}_s^- + \omega_s^2\,\Phi_s^- = i\,\frac{ke\omega_s^2}{m\omega_o^2}\,\Phi_\ell^+\,E_p^-$$

$$\ddot{\Phi}_\ell^+ + 2\Gamma_\ell\,\Phi_\ell^+ + \omega_\ell^2\,\Phi_\ell^+ = -i\,\frac{ke}{m}\,\frac{k_d^2}{k^2}\,\Phi_s^-\,E_p^+ \quad (14)$$

This is the well known result of Nishikawa.[2] (2) In this case we write $\underline{E}_p = 2E_o(\hat{x}\cos\omega_o t - \hat{y}\sin\omega_o t)$ and $\underline{k} = k_\perp\hat{x} + k_\parallel\hat{z}$ and substitution of these in Eqs. (1) and (11) yields

$$\underline{k}\cdot\delta\underline{P}_s^-(k,t) = -\frac{1}{4\pi}\,\frac{e}{m}\,\frac{k^2}{\omega_o^2}\,\chi_i(\omega_s)\,k_\perp E_o\,\Phi_\ell^+$$

$$\underline{k}\cdot\delta\underline{P}_\ell^+(\underline{k},t) = \frac{1}{4\pi}\,\frac{e}{m}\,\frac{k^2}{\omega_o^2}\,\left[1+\chi_e(\omega_s)\right]\,k_\perp E_o\,\Phi_s^- \quad (15)$$

where we have neglected the contribution from ion terms, and

$$\chi_\alpha(\omega,\underline{k}) = \sum_{n=-\infty}^{+\infty}\frac{\omega_{p\alpha}^2}{k^2}\int_{-\infty}^{+\infty}dv_\parallel\int_0^\infty v_\perp dv_\perp\left(J_n\left(\frac{k_\perp v_\perp}{\Omega_\alpha}\right)\right)^2\frac{n\Omega_\alpha}{V_\perp}\,\frac{\partial f_\alpha^o}{\partial V_\perp} +$$

$$+ k_\parallel\,\frac{\partial f_\alpha^o}{\partial V_\parallel}\Big)\Big/(\omega - n\Omega_\alpha - k_\parallel v_\parallel + i0)$$

substituting Eq. (15) in Eqs. (12) and (13). We have

$$\ddot{\Phi}_s^- + 2\Gamma_s \dot{\Phi}_s^- + \omega_s^2 \Phi_s^- = \frac{-iM\omega_s^4(\omega_o^2 - \Omega_e^2)}{8\pi n_o e^2 \omega_o^2} \chi_i(\omega_s) \frac{e}{m} k_\perp \frac{2E_o}{(\omega_o^2 - \Omega_e^2)} \Phi_\ell^+$$

$$\ddot{\Phi}_\ell^+ + 2\Gamma_\ell \dot{\Phi}_\ell^+ + \omega_\ell^2 \Phi_\ell^+ = \frac{-im(\omega_\ell^4 + \omega_\ell'^2 \Omega_e^2)(\omega_o^2 - \Omega_e^2}{8\pi n_o e^2 \omega_o^2} \left[1 + \chi_e(\omega_s)\right] \frac{e}{m} k_\perp \frac{2E_o}{(\omega_o^2 - \Omega_e^2)} \Phi_s^- \tag{16}$$

Since $\frac{\partial \varepsilon_1}{\partial \omega}\Big|_{\omega_s} = 2\frac{\omega_{pi}^2}{\omega_s^3}$ and $\frac{\partial \varepsilon_2}{\partial \omega}\Big|_{\omega_\ell} = \frac{2\omega_\ell \omega_{pe}^2}{(\omega_\ell^2 - \Omega_e^2)^2}$

the threshold field can be calculated from the following:

$$4\Gamma_s \Gamma_\ell \omega_s \omega_\ell = -\frac{\omega_s^4}{2\omega_{pi}^2} \frac{(\omega_o^2 - \Omega_e^2)^2}{2\omega_{pe}^2} \chi_i(\omega_s) \left[1 + \chi_e(\omega_s)\right] \left(\frac{e}{m}\right)^2 k_\perp^2 \frac{2E_o}{\omega_o^2 - \Omega_e^2}^2 \left(1 + \frac{\Omega_e^2}{\omega_o^2}\right)$$

$$= -\frac{\omega_s \omega_\ell}{\frac{\partial \varepsilon_1}{\partial \omega}\Big|_{\omega_s} \frac{\partial \varepsilon_2}{\partial \omega}\Big|_{\omega_\ell}} \chi_i(\omega_s) \left[1 + \chi_e(\omega_s)\right] \mu^2 \tag{17}$$

This is in agreement with Porkolab.[1,4] It is interesting to point out that if for case (2) a linearly polarized pump is used with UH and LH as decay waves, the threshold field is reduced[6,7] by a large factor $(\omega_\ell/\omega_s)^{1/2}$!

Air Force Office of Scientific Research
AFOSR-74-2668

B. R. Cheo and S. P. Kuo

REFERENCES

1. M. Porkolab, Physica, 82C, pp. 86-110 (1976).
2. K. Nishikawa, J. Phys. Soc. Japan, 24, pp. 916 and 1152 (1968).
3. Y. M. Aliev, V. P. Silin and C. Watson, Zh. Eksp. Teor. Fiz. 50, p. 943 (1966) [Sov. Phys. JETP, 23, 626 (1966)].
4. M. Porkolab, Nuclear Fusion, 12, p. 329 (1972).
5. S. Ichimaru, Basic Principles of Plasma Physics, Ch. 3, 5, (Benjamin, Inc., 1973).
6. S. P. Kuo and B. R. Cheo, Report No. EE/EP-77-027, Polytech. Inst. New York (1977).
7. S. P. Kuo, "Studies of Parametric Decay Instabilities in Magneto Plasmas," Ph. D. Dissertation, Polytech. Inst. New York (1977).

EVOLUTION OF TRANSVERSE INSTABILITY IN A HOLLOW CYLINDRICAL WEAKLY IONIZED PLASMA COLUMN

H. Kudyan and K. Chung

The formation of striations in our electrodeless ECRH device was reported previously,[1] and the possibility was pointed out that a plasma instability rather than the mechanism of electrodes might be the cause of the filamentary structures.[2] The experiment was performed on the setup described previously.[3]

Cur investigations of the background plasma of various types of plasma columns indicate that the plasma generated in our machine has significant flow fields both in the azimuthal and axial directions, and that hollow cylindrical plasma columns would develop into striations. Invariably, before breaking up into striations, strong localized oscillations which propagate azimuthally with mode numbers m = 3 and 4 are observed. The frequency of these oscillations (20 to 100 KHz) varies with the magnitude of the radial electric field but is virtually independent of ion masses. Although apparent characteristics of these oscillations are similar to edge oscillations (transverse Kelvin-Helmholty instability) observed in cylindrical alkali plasmas[4] and $\underline{E} \times \underline{B}$ rotational velocity plays a major role, our plasma is of hollow cylindrical shape and synchronous effects of surface charge waves are important in unstabilizing the hollow plasma column. As the oscillations become stronger, a distorted steady state is reached (i.e., striated plasma column).

Based on our findings from the background measurements, a small-signal model is developed which accounts for the observed transverse instability. The dispersion relation obtained in this manner is given by Eq. (1), and plotted in Figure 1. This

$$4\Omega^2 - 2\Omega\left[m\left(1 - \frac{b^2}{c^2}\right) + \left(\frac{c}{d}\right)^{2m} - \left(\frac{b}{d}\right)^{2m}\right] - \left[\left(1 - \left(\frac{c}{d}\right)^{2m}\right)\left(1 - \left(\frac{b}{c}\right)^{2m}\right) - m\left(1 - \frac{b^2}{c^2}\right)\left(1 - \left(\frac{b}{d}\right)^{2m}\right)\right] = 0 \qquad (1)$$

$$\Omega = \omega / \omega_o \quad , \quad \omega_o = e(N_e - N_i)/\varepsilon B$$

is similar to that for the diocotron instability[5] which is well-known in the case of unneutralized thin electron beams in an axial magnetic field. The roots of the dispersion relation are independent of the neutral pressure because our model neglects collisional viscosity. Experimentally, the unstable oscillations persist up to the micron range of neutral pressures. Nevertheless, the measured values of both frequency and growth rate exhibit a barely noticeable decrease with increasing neutral pressure.

The instability has a distinct effect on the configuration in our plasma. This is manifested as the formation of striations. An understanding of the phenomena can be

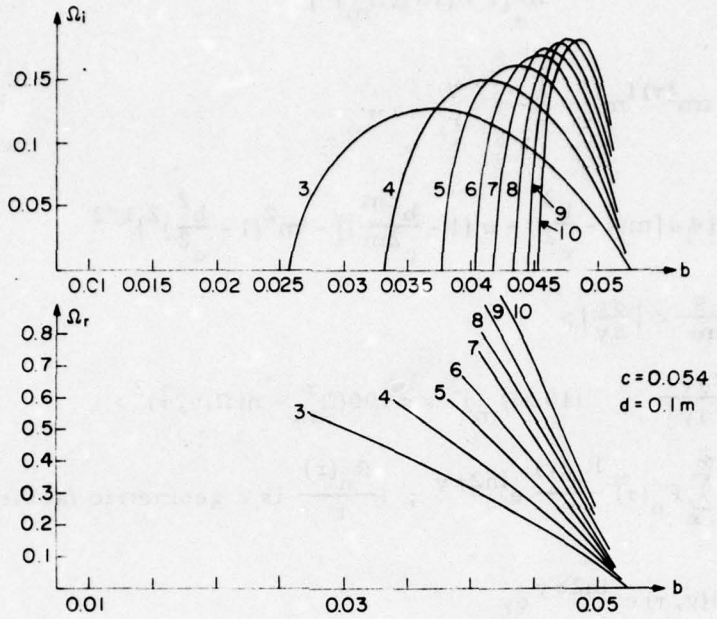


Fig. 1. The real and positive imaginary parts of the roots of the dispersion relation for m = 3 to 10 as a function of b (c = 0.05 m, d = 0.1 m).

formulated by treating the hollow column plasma to be composed of two individual systems: the boundaries which can support unstable surface charge waves, and the bulk of the plasma. The boundary system is taken to be always uniform in the azimuthal direction while the internal layers of charge are assumed to be always uniform in the radial direction. These two "systems" communicate solely via the electrostatic fields. As the fields of the two interacting surface charge waves grows, the internal layers of charge (which are in synchronism with the imposed wave field due to their local velocity) respond to it by bunching. The bunching in turn transmits a field which is not in phase with the imposed fields and makes the two surface waves progressively asynchronous. Eventually, the two interacting surface waves lose syncrhonism and become marginally stable. The bunching itself cannot go on indefinitely because the internal layers "see" a progressively increasing average Doppler frequency. Eventually evolution halts and a distorted steady state is reached.

The model explained above can be represented by a two dimensional mathematical model, which a priori assumes the presence of the field generated by the surface charge waves (a consequence of the radial inhomogeneity). The set of nonlinear equations scaled to one linear growth time and one perimeter of synchronous layer is:

$$\frac{\partial f}{\partial \tau} + \frac{5}{\pi}\Omega(y,\tau)\frac{\partial f}{\partial y} = \frac{\partial}{\partial y}\sum_m \frac{\zeta(y)}{[1+(10\Delta\Omega_m)^2]} I_m(\tau)\frac{\partial f}{\partial y} \tag{2}$$

$$\frac{\partial I_m}{\partial \tau} = 20\,\Omega_{im}(\tau) I_m \tag{3}$$

where

$$\Omega_{im} \simeq +\frac{1}{4}\{4\alpha[m(1-\frac{b^2}{c^2}) - \alpha(1-\frac{b^{2m}}{c^{2m}})] - m^2(1-\frac{b^2}{c^2})^2\}^{1/2}$$

$$\alpha(\tau) = 1 + \frac{5}{m\pi} < \left|\frac{\partial f}{\partial y}\right| >$$

$$\zeta(y) = \frac{\partial f/\partial y}{|\partial f/\partial y|} \quad , \quad (10\Delta\Omega_m)^2 = <100(\Omega_m^r - m\Omega(y,\tau)^2>$$

$$\Omega(y,\tau) = \frac{1}{2}\sum_{-\infty}^{\infty} F_n(\tau)\frac{R_n(r)}{r} e^{in2\pi y} \quad , \quad (\frac{R_n(r)}{r} \text{ is a geometric factor})$$

$$F_n(\tau) = \int_o^1 f(y,\tau)\, e^{-in2\pi y}\, dy$$

$f(y,\tau)$ is the normalized charge density and

$I_m(\tau)$ is the normalized electric field intensity of each unstable surface wave.

These equations are followed in time numerically using the IGI (Interactive Graphic Interpretive) language. The outcome of these computations indicate that the initially uniform configuration evolves in time, as seen in Figs. 2 and 3, toward a distorted steady state. In our experimental investigations of the conditions leading to striations, we have noticed that with very thin hollow columns (i.e., several unstable surface modes) it was usually hard to induce uniformly distributed striations around the column perimeter; instead a single or several striations were noticed to exist asymmetrically. With relatively thicker hollow columns (i.e., one unstable surface mode), the striations were noticed to occur rather symmetrically around the column perimeter. Our nonlinear model does exhibit this characteristic behavior of the striations. Figure 2 corresponds to an initial column geometry where only the m = 3 surface mode is unstable; Fig. 3 corresponds to an initial column geometry where m = 3 and 4 surface modes are simultaneously unstable.

Joint Services Technical Advisory Committee
F44620-74-C-0056

H. Kudyan and K. Chung

REFERENCES

1. K. Chung, Bull. Am. Phys. Soc., Vol. 14, p. 830 (1969).
2. K. Chung, J. Appl. Phys., Vol. 40, p. 3885 (1969).
3. H.M. Kudyan and K. Chung, "A Kelvin-Helmholtz Type Instability in Plasma Striations," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76, pp. 213-215 (1976).
4. G.I. Kent, N.C. Jen and F. Chen, Phys. Fluids, Vol. 12, p. 2140 (1969).
5. P. Gwenard and H. Huber, Ann. Radioel., Vol. 7, p. 252 (1952).

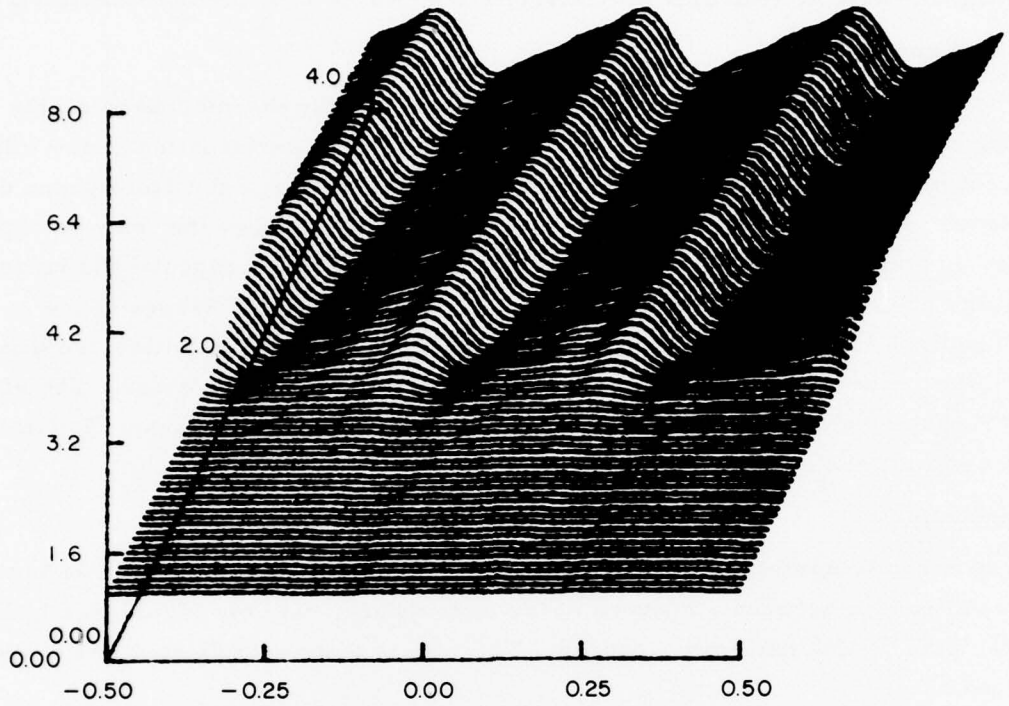


Fig. 2. The computational result for the case where only m = 3 is unstable.

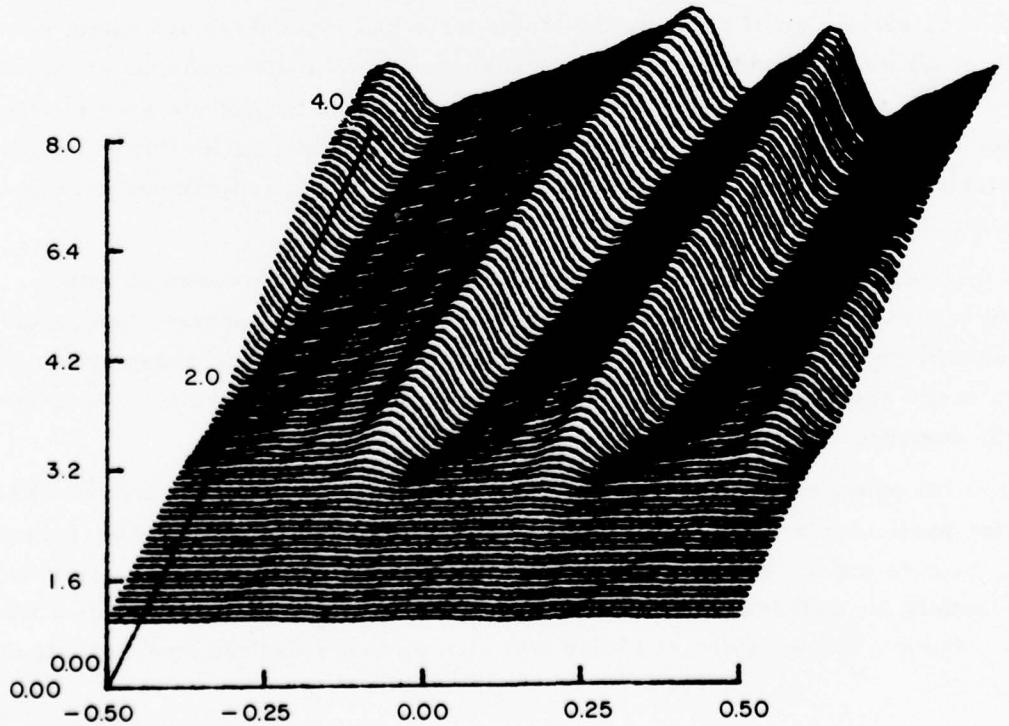


Fig. 3. The computational result for the case where m = 3 and m = 4 are unstable.

RESPONSE OF MINOR SPECIES TO GRAVITY WAVES IN THE THERMOSPHERE

S. H. Gross and H. Eun

The response of minor species to gravity waves in the thermosphere varies according to the mass of the species. The relative density perturbation of any minor constituent may be related to the relative density perturbation of the atmosphere due to the waves via a complex function of frequency and wave number that may be represented as an amplitude and phase response. Peaks and dips in response and large phase shifts are found that are associated with complex poles and zeroes of the response function. These poles and zeroes depend on background quantities, so that the nature of the response is model dependent; collisions are most important. The effects of background and collisions are examined using numerical computations. Relationship to some satellite observations are discussed.

A. Introduction

It is recently appreciated that gravity waves may play an important role in the transport of energy between all layers of the atmosphere, not only for the earth, but for all the planets that have atmospheres. The role is quite complicated and not fully understood as yet.

One aspect of this problem is the effect of gravity waves on the various constituents, not least of which is ionization because of detectability. These disturbances may influence refraction of radio by the atmosphere and ionosphere and affect communications as well as radio scientific measurements. The disturbances of minor species may also influence their chemical reactions, so as to produce special effects of greater consequences as, for example, nitrogen oxides in relationship to ozone. These possibilities deserve further investigation, and they are mentioned here in order to illustrate the relevance of the topic of this report.

In this report we are concerned with the effects of gravity waves on minor neutral species that are in the thermosphere. These individual species are recently measurable by satellite. Their response is quite complex because of dependence of the mass of the species relative to the mean mass of the atmosphere at a given level and on the dynamic background motion.

Here the phase and amplitude response relative to the main atmosphere will be treated for particular wave frequencies and horizontal wave numbers for various minor species, such as argon, helium and molecular oxygen. Examples will be given for various models for different times of the day, different solar activity and for a typical altitude, 250km. The behavior at 170km will also be described for contrast. It will

be shown that large phase shifts and large relative amplitude responses are possible at 250km altitude for specific frequencies and wave numbers for some species. It will also be shown that the response is highly dependent on the magnitude of the collision frequency between species. Major amphasis will be on long period (~ 2 hours) waves of long horizontal wavelength (~ 1000's km) in order to compare with satellite measurements.

Data from two satellites are of interest. Reber et al.[1] found from measurements on board the Atmospheric Explorer Satellite, AE-C, that the helium response was consistently 180° out of phase with nitrogen and argon at altitudes from 160 km to 250 km. The response of argon and nitrogen were nearly in phase, but the amplitude response of argon was roughly four times that of helium and twice that of nitrogen. This response was evident for horizontal wavelengths between 100 km and 400 km, though much longer wavelengths were also evident in the data. Trinks and Mayr[2] found evidence of similar responses in the data from the European Satellite ESRO 4. At about 240 km altitude behavior was found to be about 171° out of phase with nitrogen and argon which were in phase. The amplitude of the argon response was found to be about 2.7 times that of helium and about 1.6 times that of nitrogen. The horizontal wavelength and period of these waves were found to be 3000 km and 150 minutes, respectively. It is with such waves that we wish to make theoretical comparisons.

B. Theory

The basic theory was developed by Gross and Eun[3] for two fluids, one major and one minor in local thermodynamic equilibrium. The medium is assumed to be plane stratified and the minor species diffuses in the major according to its background distribution. The perturbation equations are given in matrix form as follows:

$$\frac{de}{dz} = \hat{T}e \quad , \quad \text{where } e = \begin{pmatrix} c_{1z} \\ p_1/p_o \\ C_{alz} \\ p_{al}/p_{ao} \end{pmatrix}$$

$$\hat{T} = \begin{pmatrix} T_{11} & T_{12} & 0 & 0 \\ T_{21} & T_{22} & 0 & 0 \\ T_{31} & T_{32} & T_{33} & T_{34} \\ T_{41} & T_{42} & T_{43} & T_{44} \end{pmatrix} \qquad (1)$$

Here e is a column matrix of 4 elements, c_{1z} is the vertical component of the perturbation velocity of the main atmosphere, p_1/p_o is the relative pressure

perturbation of the main atmosphere, C_{a1z} is the vertical component of the perturbation peculiar velocity of the minor species and p_{a1}/p_{ao} is the relative pressure perturbation of the minor species. Subscript o refers to background quantities, whereas subscript 1 refers to perturbation quantities. Perturbation is caused by a gravity wave moving through the atmosphere. The vertical direction is denoted by the coordinate z. The $\hat{T}$ matrix is a 4 x 4 matrix with elements T_{ij}, where i = 1 to 4, j = 1 to 4 as shown in Equation (1). Matrix elements for an isothermal atmosphere in diffusive equilibrium are as follows:

$$T_{11} = \frac{1}{\gamma_n H_o}, \quad T_{12} = \frac{i\omega}{\gamma_n}\left(1 - k_x^2 \frac{a_{no}^2}{\omega^2}\right), \quad T_{13} = T_{14} = 0$$

$$T_{21} = \frac{i\omega\gamma_n}{a_{no}^2}\left(1 - \frac{N_n^2}{\omega^2}\right), \quad T_{22} = \frac{\gamma_n - 1}{\gamma_n H_o}, \quad T_{23} = T_{24} = 0$$

$$T_{31} = \frac{1}{H_a} - \frac{1}{H_o}, \quad T_{32} = -i\omega + \frac{k_x^2 a_{no}^2}{\gamma_n(-i\omega + \nu_{ano})}, \quad T_{33} = \frac{1}{H_a}, \tag{2}$$

$$T_{34} = i\omega - \frac{k_x^2}{-i\omega + \nu_{ano}} \frac{p_{ao}}{\rho_{ao}}, \quad T_{41} = i\omega\left(1 - \frac{N_n^2}{\omega^2}\right) \frac{\rho_{ao}}{p_{ao}},$$

$$T_{42} = \frac{g(\gamma_n - 1)}{\gamma_n} \frac{\rho_{ao}}{p_{ao}}, \quad T_{43} = (i\omega - \nu_{ano}) \frac{\rho_{ao}}{p_{ao}}, \quad T_{44} = 0.$$

Here γ_n is the ratio of specific heats, H_o is the scale height of the main atmosphere, ω is the radian frequency, k_x is the horizontal component of the wavenumber, a_{no} is the velocity of sound in the major atmospheric constituent, N_n is the buoyancy or Brunt-Värsälä frequency, H_a is the scale height of the minor species, ν_{ano} is the collision frequency of a minor species particle with the major constituent, p_{ao} and ρ_{ao} are the background pressure and density, respectively, of the minor species particles, and g is the acceleration of gravity.

On substituting $\frac{d}{dz} = ik_z$ in Eq. (1), where k_z is the vertical component of the wavenumber, one may solve for the ratio p_{a1}/p_{ao} in terms of p_1/p_o as a function of ω, k_x and k_z. k_z may then be eliminated on using the dispersion relationship for gravity waves which relates k_z, ω and k_x. The ratio (p_{a1}/p_{ao}) to (p_1/p_o) is then the relative pressure response of the minor species wave to the pressure perturbation of the main atmosphere. This response is a complex function of ω and k_x, and, one may determine a phase and amplitude from this function. Though the above yields

expressions for the relative pressures, this procedure was followed for the relative densities, preferable because density is actually measured. The details of the differences in the matrix elements for these densities as contrasted with the pressures, however, will not be given here, but will be published elsewhere.

One may, nevertheless, illustrate the response under simplifying conditions, as shown in Table I. For the purpose here it is based on the assumption that the frequency $\omega << N_n$. There are several solutions depending on the magnitude of the real part of k_z, designated k_{zr}, relative to the inverse scale height as well as the magnitude of the collision frequency (low altitude - high collision rate) high altitude - low collision rate). It is seen in three out of the four cases in the Table that the response depends on the mass ratio of the minor to major species particles m_a/m_o. Large phase shift is possible only for light weight minor particles, in accordance with satellite experimental findings.

TABLE I.

$$\frac{\rho_{a1}}{\rho_{ao}} = A(\omega, k_x, k_{zr})\, e^{j\varphi(\omega, k_x k_{zr})}\, \frac{\rho_1}{\rho_o}$$

Low Frequency Approximation, $\omega << N_n$

Range of k_{zr}	Range of ν_a	Solution of $\frac{\rho_{a1}}{\rho_{ao}}$	Relative Phase to the Main Disturbance
$k_{zr} >> \frac{1}{H_o}, \frac{1}{H_a}$	$k_{zr} >> \frac{\omega \nu_a}{C_a^2}$ (high altitude)	$\frac{\rho_1}{\rho_o}$	In Phase
	$k_{zr} << \frac{\omega \nu_a}{C_a^2}$ (low altitude)	$\frac{\rho_1}{\rho_o} \frac{(\gamma \frac{m_a}{m_o} - 1)}{(\gamma - 1)}$	$\frac{m_a}{m_o} < \frac{1}{\gamma}$; 180° out of phase $\frac{m_a}{m_o} > \frac{1}{\gamma}$; In Phase
$k_{zr} << \frac{1}{H_o}, \frac{1}{H_a}$ (above 200 km)	$k_{zr} >> \frac{\omega \nu_a}{C_a^2}$	$-\frac{\rho_1}{\rho_o}(1 - 2\frac{m_a}{m_o})$	$\frac{m_a}{m_o} < \frac{1}{2}$; 180° out of phase $\frac{m_a}{m_o} > \frac{1}{2}$; In Phase
	$k_{zr} << \frac{\omega \nu_a}{C_a^2}$	$-\frac{\rho_1}{\rho_o} \frac{1}{(\gamma - 1)}[(2-\gamma)\frac{m_a}{m_o} - 1]$	In phase for species with $\frac{m_a}{m_o} < 3$

For more complex cases one must resort to numerical analysis. This was done for a series of background models ranging from low to high solar activity. A typical case is shown in Fig. 1 which is for mean solar activity, at midnight, 250 km altitude, 150 minute period wave. The plot is against the horizontal wavelength λ_x. It shows phase and amplitude for helium (He), oxygen (O_2) and argon (Ar). It should be noted that all species have relative amplitudes greater than unity in this range. The large phase shift of helium is quite evident, whereas there is little phase shift between oxygen and argon.

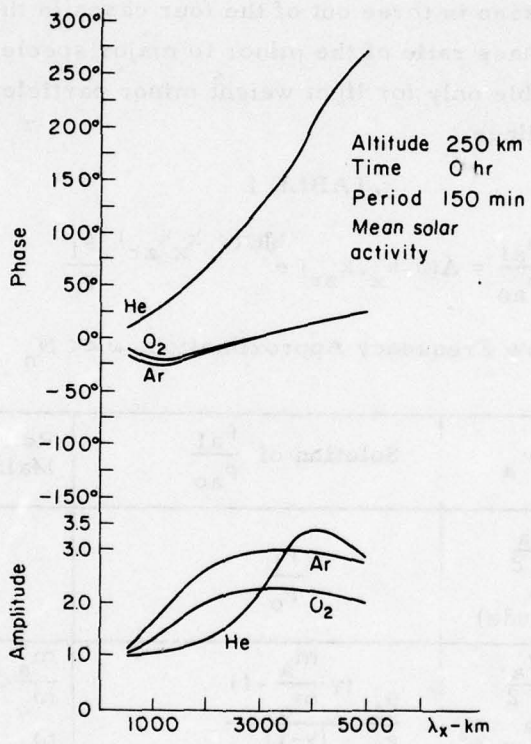


Fig. 1. Density phase and amplitude relative to main disturbance.

As solar activity is reduced at 250 km altitude for this period, it is found that the argon and oxygen responses hardly change, whereas the helium response becomes peaked at specific wavelengths. Similar characteristics are found for other periods, as well.

The results are summarized in Fig. 2 where the control of collision frequency is evident. This figure is a plot of the peak of the helium response as a function of collision frequency. The various models refer to CIRA 1965 atmospheric models.[4]

The higher is the number of the model, the higher the solar activity. The graph illustrates the peak response of helium for three wave periods, 130 minutes, 150 minutes and 180 minutes. The corresponding horizontal wavelengths for the various points are indicated by number in the body of the figure.

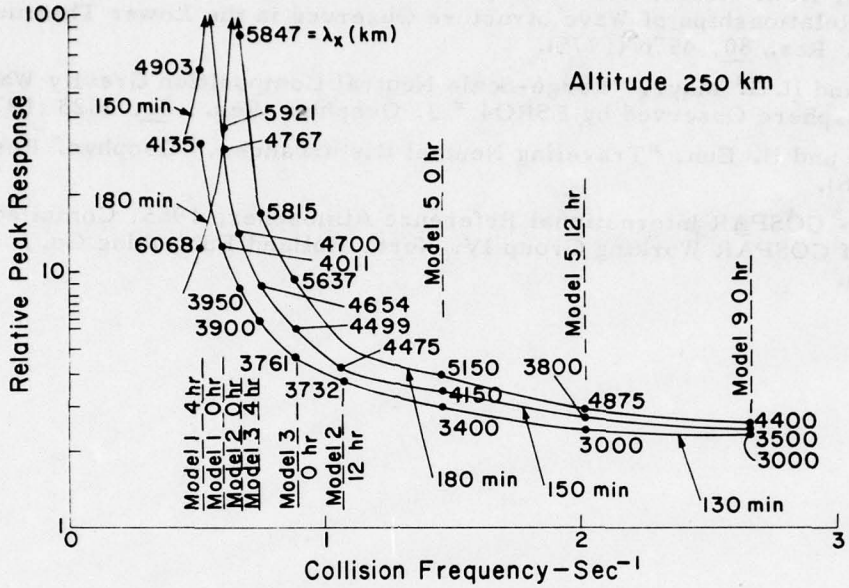


Fig. 2. Relative peak response of helium vs. helium - atomic oxygen collision frequency for wave periods of 180, 150 and 130 min.

Calculations made for an altitude of 170 km show that the phase and amplitude responses are nearly model independent. Nevertheless the phase shifts of helium are quite large. The collision frequencies at this altitude are about an order of magnitude greater than that at 250 km, and the absence of sensitivity to solar activity, may be explained on this basis.

In conclusion, calculations have shown that the thermosphere is capable of a strong response for minor species of lighter mass at the proper range of altutudes. It also shows that the calculated responses are very much like that found from satellite measurements for the solar activity at the time. Typically, strong response is for periods ~ 2 hours. Helium tends to have a very strong resonant response in the 200 - 300 km altitude range for lower solar activity, suggesting the likelihood of coupling to its acoustic wave.[3] At higher solar activity the argon response grows to dominate, though helium still exhibits large phase shift. The response of argon is mainly independent of atmospheric models. The relative response of minor species may possible be utilized in the future as an indicator of local heating effects because of the sensitivity, particularly for helium, to collision rates.

National Aeronautics and Space Administration
NGR33-006-047

S. H. Gross and H. Eun

REFERENCES

1. C. A. Reber, A. E. Hedin, D. T. Pelz, W. E. Potter and L. H. Brace, "Phase and Amplitude Relationships of Wave Structure Observed in the Lower Thermosphere," J. Geophys. Res., 80, 4576 (1975).

2. H. Trinks and H. G. Mayr, "Large-Scale Neutral Composition Gravity Waves in the Thermosphere Observed by ESRO4," J. Geophys. Res., 81, 4023 (1976).

3. S. H. Gross and H. Eun, "Traveling Neutral Disturbances," Geophys. Res. Lett., 3, 257 (1976).

4. CIRA 1965 - COSPAR International Reference Atmosphere 1965, Compiled by the Members of COSPAR Working Group IV, North-Holland Publishing Co., Amsterdam.

NEW RESULTS IN ELECTRICAL MACHINE THEORY: A FLOQUET-THEORY OF THE GENERAL, LINEAR NON-IDEAL ROTATING MACHINE

D.C. Youla and J.J. Bongiorno, Jr.

A. Introduction

Conceptually, an electromagnetic rotating machine is a collection of n windings threaded either in air or around ferromagnetic material or both which are disposed into two groups, the stator and the rotor. The former contains s windings and is stationary while the latter contains the remaining n-s=r windings and rotates as a rigid body about an axis fixed in the laboratory frame. If eddy currents, hysteretic effects and retardation are negligible it can be assumed that the instantaneous flux linking each winding at time t is a function of the n winding currents $i_1(t), i_2(t), \cdots, i_n(t)$ and a single angle $\theta(t)$ which serves to define the electro-geometric configuration. Thus, over a suitable dynamic range the flux

$$\Psi_k = \Psi_k[i_1(t), i_2(t), \cdots, i_n(t); \theta(t)] \tag{1}$$

linking winding #k is a single-valued function of the n+1 arguments $i_1, i_2, \cdots, i_n$, θ with period 2π in θ, $k=1\rightarrow n$. Ordinarily, the angle θ is related to true mechanical degrees ϕ via a formula of generic type

$$\theta = \frac{p}{2} \cdot \phi \quad , \tag{2}$$

p the number of poles.

Let the current $i_k(t)$ be supplied by the voltage source $v_k(t)$ and denote the mutual winding resistance by $r_{k\ell}, (k, \ell = 1 \rightarrow n)$. Then, by applying Kirchoff's voltage law to each winding we obtain the equations

$$v_k(t) = \sum_{\ell=1}^{n} r_{k\ell} i_\ell(t) + \frac{d\Psi_k(t)}{dt} \quad , \tag{3}$$

$k=1\rightarrow n$; or, more compactly,*

$$v(t) = Ri(t) + \dot{\Psi}[i(t), \theta(t)] \tag{4}$$

where

$$\begin{aligned} v &= (v_1, v_2, \cdots, v_n)' \quad , \\ i &= (i_1, i_2, \cdots, i_n)' \quad , \\ \Psi &= (\Psi_1, \Psi_2, \cdots, \Psi_n)' \end{aligned} \tag{5}$$

---

* $\dot{a}(t) \equiv \frac{da(t)}{dt}$

are column-vectors ($'$ denotes matrix transpose) and

$$R(\theta) = (r_{k\ell}(\theta)) \tag{6}$$

is the associated nxn resistance matrix. Omitting the argument t,

$$v = R(\theta)i + \dot{\Psi}(i,\theta) \quad , \tag{7}$$

the preferred working form of Equation (3). As is well known, in a linear machine flux and current satisfy the linear relationship

$$\Psi(i,\theta) = L(\theta)i \quad , \tag{8}$$

in which $L(\theta)$ is the nxn inductance matrix. In general,[1] $R(\theta)$ and $L(\theta)$ are real, of period 2π in θ and symmetric nonnegative and positive-definite, respectively; i.e., $R(\theta) = R(\theta+2\pi)$, $L(\theta) = L(\theta+2\pi)$, $R'(\theta) = R(\theta) \geq 0$ and $L'(\theta) = L(\theta) > 0$ for all θ.

The substitution $i = L^{-1}\Psi$ converts Eq. (7) into the first-order linear vector differential equation for flux,

$$\dot{\Psi} = -A(\theta)\Psi + v \tag{9}$$

whose defining coefficient matrix

$$A(\theta) = R(\theta)L^{-1}(\theta) \tag{10}$$

is evidently of period 2π. However, because A is a function of θ instead of $\dot{\theta}$, this description is time-variable even under the condition of constant angular velocity. Unfortunately, if the machine is not assumed to be ideal in the sense of Park (Appendix), this lack of stationarity cannot be transformed away by any obvious change of reference frame. Nevertheless, such a frame, the Floquet frame does exist.

B. The Floquet Frame

Let us imagine that we attach to the rotor n open-circuited exploratory coils whose various coefficients of coupling are permitted to depend explicitly on both $\theta(t)$ and $\dot{\theta}(t)$. Then, it is reasonable to assume that the flux vector $\Psi_N(t)$ sampled by these coils has the form

$$\Psi_N(t) = T(\theta(t), \dot{\theta}(t))\Psi(t) \tag{11}$$

where the nxn matrix $T(\theta, \mu)$ possesses the following properties:

P_1. For every fixed $\mu > 0$, $T(\theta, \mu)$ is real, nonsingular and of period 2π in θ.

P_2. $T(\theta, \mu)$ has continuous partial derivatives with respect to θ and μ. Any matrix $T(\theta, \mu)$ satisfying P_1 and P_2 is said to be admissible and defines an admissible reference frame.

Using Eq. (11) to eliminate Ψ in Eq. (9) we easily obtain the differential equation for $\Psi_N(t)$,

$$\dot{\Psi}_N = (\dot{\theta}\, T_\theta T^{-1} - TAT^{-1} + \ddot{\theta}\, T_\mu T^{-1})\Psi_N + v_N \quad . \tag{12}$$

In Eq. (12),

$$v_N = v_N(t) = T(\theta(t), \dot{\theta}(t))v(t) \quad , \tag{13}$$

$$T_\theta \equiv \frac{\partial T(\theta(t), \dot{\theta}(t))}{\partial \theta} \tag{14}$$

and

$$T_\mu \equiv \frac{\partial T(\theta(t), \dot{\theta}(t))}{\partial \mu} \quad . \tag{15}$$

($\Psi_N(t), v_N(t)$, etc., are "normalized" vectors.)

According to a result of Erugin,[2] for all sufficiently large μ there exists an admissible matrix $T(\theta, \mu)$ such that

$$T(\theta, \mu)A(\theta)T^{-1}(\theta, \mu) - \mu T_\theta(\theta, \mu)T^{-1}(\theta, \mu) = \mu C(\mu) \tag{16}$$

where $C(\mu)$ is real and independent of θ. *In fact T and C admit* Laurent expansions in $1/\mu$ valid in some neighborhood of $\mu = \infty$. More explicitly,

$$T(\theta, \mu) = \sum_{k=0}^{\infty} T_k(\theta)\mu^{-k} \tag{17}$$

and

$$C(\mu) = \sum_{k=0}^{\infty} C_k \mu^{-k} \tag{18}$$

where (1 is the identity and 0 is the zero matrix),

F_1) all $T_k(\theta)$ are real, nxn, independent of μ and of period 2π in θ, $k=0 \rightarrow \infty$.

F_2) $T_0(0) = 1$ and $T_k(0) = 0$, $k=1 \rightarrow \infty$.

F_3) All C_k are real, nxn and constant, $k=0 \rightarrow \infty$.

Under such a choice for $T(\theta, \mu)$, Eq. (12) collapses into the simpler form

$$\dot{\Psi}_N = -(\dot{\theta}C(\dot{\theta}) - \ddot{\theta}T_\mu T^{-1})\Psi_N + v_N \quad . \tag{19}$$

In particular, *for constant rotor velocity*, $\ddot{\theta}(t) \equiv 0$ and Eq. (19) reduces to

$$\dot{\Psi}_N = -\dot{\theta}C(\dot{\theta})\Psi_N + v_N \quad , \tag{20}$$

a constant-coefficient differential equation. We have therefore achieved out desired objective and established a Park-like circuit description for the general linear rotating machine which is valid in some Floquet frame characterized by the pair of Floquet matrix parameters $T(\theta,\mu), C(\mu)$.

There exists a class of machines, known as ideal, for which $T(\theta,\mu)$ can be chosen independent of μ. For this class, which includes the Park machine (Appendix),

$$A(\theta) = e^{C_0\theta} C_1 e^{-C_0\theta} \quad , \tag{21}$$

$$T(\theta,\mu) = T_0(\theta) = e^{-C_0\theta} \tag{22}$$

and

$$C(\mu) = C_0 + C_1/\mu \tag{23}$$

where C_0 and C_1 are two real nxn matrices.* Clearly, for such a machine, the $\ddot{\theta}$ term in Eq. (19) vanishes rigorously and the normalized differential equation reduces to

$$\dot{\Psi}_N = -(\dot{\theta}C_0 + C_1)\Psi_N + v_N \quad . \tag{24}$$

The question of how large μ should actually be to assure the convergence of the expansions in Eqs. (17) and (18) is a difficult one to answer quantitatively. Nevertheless, since μ is always identified with $\dot{\theta}(t) = \omega = 2\pi f \approx 377$ radians/sec. for 60 Hz machines, $\mu \gg 1$ and we conjecture that Eqs. (17) and (18) converge in most practical cases.

Accepting the truth of the above hypothesis, it is easily seen that under the same conditions and in any suitable norm,

$$\| \ddot{\theta} T_\mu(\theta,\dot{\theta}) T^{-1}(\theta,\dot{\theta}) \| \sim \frac{|\ddot{\theta}|}{(\dot{\theta})^2} = \frac{|\dot{\omega}|}{\omega^2} \tag{25}$$

is a compatible estimate. Consequently, because of the minuteness of the ratio $|\dot{\omega}|/\omega^2$, it appears that the acceleration term in Eq. (19) should also be negligible in practical machines. Thus, we have supplied a plausible physico-mathematical basis for the following extrapolation.

The Practical-Machine Theorem. For practical synchronous machines there exists a suitable Floquet frame in which the circuit description of the machine assumes the extremely accurate acceleration-independent form,

$$\dot{\Psi}_N = -\dot{\theta}C(\dot{\theta})\Psi_N + v_N \quad . \tag{26}$$

This is a linear differential equation whose coefficient matrix depends solely on $\dot{\theta}(t)$.

---

*Of course, $A(\theta)$ must have period 2π. Incidentally, for the Park machine C_0 is skew-symmetric and $T_0(\theta)$ is orthogonal (Appendix).

C. An Experimental Plus Theoretical Overview

According to Eqs. (17) and (18),

$$\mu C(\mu) = \mu C_0 + C_1 + \text{terms of order } (\frac{1}{\mu}) \tag{27}$$

and

$$T(\theta, \mu) = T_0(\theta) + \frac{T_1(\theta)}{\mu} + \text{terms of order } (\frac{1}{\mu^2}) \quad . \tag{28}$$

Hence, for $\dot{\theta}(t) \to \infty$,

$$\dot{\theta} C(\dot{\theta}) \sim \dot{\theta} C_0 + C_1 \quad , \tag{29}$$

$$T(\theta, \dot{\theta}) \sim T_0(\theta) \quad . \tag{30}$$

In other words, as $\dot{\theta} = \omega \to \infty$, every machine behaves like an ideal machine whose $A(\theta)$ matrix is given by

$$A_\infty(\theta) = e^{C_0 \theta} C_1 e^{-C_0 \theta} \quad . \tag{31}$$

This immediately suggests that there should exist experimental procedures for the measurement of C_0 and C_1.*

Equally important, however, and quite consistent with the Practical-Machine theorem, is the real possibility that the approximations

$$\dot{\theta} C(\dot{\theta}) \sim \dot{\theta} C_0 + C_1 \tag{32}$$

and

$$T(\theta, \dot{\theta}) \sim T_0(\theta) + \frac{T_1(\theta)}{\dot{\theta}} \tag{33}$$

are sufficiently accurate over the actual ranges of working frequencies of most practical machines.

Naturally, a substantiation of this conjecture must involve an in-depth study of the properties of the C_r's and T_r's and their relationship to the given machine data, $A(\theta) = R(\theta)L^{-1}(\theta)$. To give an idea of what is involved we shall derive the necessary recursions and offer a brief analysis of the attendant difficulties.

Consider the periodic differential equation

*Currently under investigation.

$$\frac{dX(\theta)}{d\theta} = \frac{1}{\mu} \cdot X(\theta) A(\theta) \tag{34}$$

and let us seek to construct a fundamental matrix solution of the form

$$X(\theta, \mu) = e^{C(\mu)\theta} T(\theta, \mu) \tag{35}$$

wherein $T(\theta, \mu)$ and $C(\mu)$ possess the expansions Eqs. (17) and (18), respectively, and the enumerated properties F_1)-F_3). The substitution of Eq. (35) into Eq. (34) yields immediately*

$$T_\theta = -CT + \frac{1}{\mu} TA \quad , \tag{36}$$

and upon equating coefficients of like powers of $1/\mu$ on both sides we obtain the coupled system of equations,

$$\frac{dT_0}{d\theta} = -C_0 T_0 \quad , \tag{37}$$

$$\frac{dT_r}{d\theta} = -\sum_{k=0}^{r} C_k T_{r-k} + T_{r-1} A, \quad r \geq 1 \quad . \tag{38}$$

The solution of Eq. (37) subject to the initial condition $T_0(0) = 1$ is obviously

$$T_0(\theta) = e^{-C_0\theta} \quad . \tag{39}$$

However, the solutions of Eq. (38) subject to the initial conditions $T_r(0) = 0$, $r=1 \rightarrow \infty$, are best expressed in normalized form. Thus, for $r=0 \rightarrow \infty$, let

$$Z_r(\theta) = T_0^{-1}(\theta) T_r(\theta) = e^{C_0\theta} T_r(\theta) \tag{40}$$

and

$$M_r(\theta) = T_0^{-1}(\theta) C_r T_0(\theta) \quad . \tag{41}$$

Then, omitting the straightforward details, we find that

$$Z_1(\theta) = \int_0^\theta (A(x) - M_1(x))dx \quad , \tag{42}$$

and for $r \geq 1$,

$$Z_r(\theta) = \int_0^\theta (Z_{r-1}(x)A(x) - M_1(x)Z_{r-1}(x) - \sum_{k \neq 0, 1, r} M_k(x)Z_{r-k}(x) - M_r(x))dx \quad . \tag{43}$$

---

*Equation (16) is a rearranged version of Equation (36).

The algorithm is initialized by first selecting the real matrix C_0 so that $T_0(\theta)$ has period 2π. Then, since the integral of a periodic function is periodic iff the function has zero average value, it follows from Eq. (42) that $T_1(\theta)$, and therefore $Z_1(\theta)$, has period 2π iff

$$\int_0^{2\pi} A(\theta)d\theta = \int_0^{2\pi} e^{C_0\theta} C_1 e^{-C_0\theta} d\theta \quad . \tag{44}$$

This average-value constraint determines an acceptable C_1 (if one exists). In general, having determined $C_0, C_1, \cdots, C_{r-1}$, we obtain C_r by applying the same average-value argument to Eq. (43):

$$\int_0^{2\pi} (Z_{r-1}(\theta)A(\theta) - M_1(\theta)Z_{r-1}(\theta) - \sum_{k\neq 0,1,r}^{r} M_k(\theta)Z_{r-k}(\theta))d\theta = \int_0^{2\pi} e^{C_0\theta} C_r e^{-C_0\theta} d\theta \quad ,$$

$$r \geq 1 \quad . \tag{45}$$

The "best" <u>physical</u> choice for C_0 is not at all apparent although the simplest mathematical choice that always works is $C_0 = 0$. Under this initialization, $T_0(\theta) = 1$,

$$\frac{1}{2\pi}\int_0^{2\pi} A(x)dx = C_1; \quad T_1(\theta) = \int_0^{\theta} A(x)dx - \theta C_1 \quad , \tag{46}$$

$$\frac{1}{2\pi}\int_0^{2\pi} (T_{r-1}(\theta)A(\theta) - C_1 T_{r-1}(\theta) - \sum_{k\neq 0,1,r}^{r} C_k T_{r-k}(\theta))d\theta = C_r \tag{47}$$

and

$$T_r(\theta) = \int_0^{\theta} (T_{r-1}(x)A(x) - C_1 T_{r-1}(x) - \sum_{k\neq 0,1,r}^{r} C_k T_{r-k}(x))dx - \theta C_r \quad ,$$

$$r \geq 1 \quad . \tag{48}$$

To illustrate the inappropriateness of the initialization $C_0 = 0$, consider the ideal machine with A matrix $A(\theta) = e^{-C_0\theta} C_1 e^{C_0\theta}$, Equation (21). Using this $C_0 (\neq 0)$ and C_1 we obtain the <u>exact</u> one and two-term expressions

$$T(\theta, \mu) = T_0(\theta) = e^{-C_0\theta} \tag{49}$$

and

$$C(\mu) = C_0 + \frac{C_1}{\mu} \quad . \tag{50}$$

But unfortunately, the initial choice $C_0 = 0$ leads, via Eqs. (46)-(48), to two infinite-series expansions in $1/\mu$ for both $T(\theta, \mu)$ and $C(\mu)$.

Our studies up to this point seem to suggest that an "optimal" choice of C_0 should probably be predicated on the minimization of some error norm,

$$\| A(\theta) - e^{C_0\theta} C_1 e^{-C_0\theta} \| = \| A(\theta) - A_\infty(\theta) \| \quad , \tag{51}$$

measuring the "distance" from $A(\theta)$ to its very-high-frequency limit, $A_\infty(\theta)$. We are confident of the correctness of this approach and any new developments will be documented in future progress reports.

D. Appendix: The Ideal Park Machine

Park's ideal 3-phase machine occupies a central place in the conventional theory and continues to play a key role in studies involving interconnected power systems.[1] For the reader's benefit and for future reference we shall give explicit formulas for all the pertinent design data and all the Floquet parameters without neglecting either damper-bar inductance or damper-bar resistance.

P_{k1}) The three identical stator phases are denoted by "a," "b," "c" and are assumed to possess the same constant ohmic resistance $r_s \geq 0$. The $\ell x \ell$, real, constant symmetric nonnegative-definite matrix R_r incorporates all field-winding loss and all damper-bar resistance. Thus,*

$$R(\theta) = R = r_s 1_3 \dotplus R_r = R' \geq 0 \quad . \tag{52}$$

P_{k2}) The spacial separation between contiguous phases is equivalent to 120 electrical degrees. Let θ denote the electrical angle relative to phase a. Then,

$$L(\theta) = \left[\begin{array}{c|c} L_{ss}(\theta) & L_{sr}(\theta) \\ \hline L'_{sr}(\theta) & L_{rr}(\theta) \end{array}\right] = L'(\theta) > 0 \tag{53}$$

where

$$L_{ss}(\theta) = \begin{bmatrix} L_{aa}(\theta) & L_{bc}(\theta + 120^\circ) & L_{bc}(\theta - 120^\circ) \\ L_{bc}(\theta + 120^\circ) & L_{aa}(\theta - 120^\circ) & L_{bc}(\theta) \\ L_{bc}(\theta - 120^\circ) & L_{bc}(\theta) & L_{aa}(\theta + 120^\circ) \end{bmatrix} \tag{54}$$

---

*$A \dotplus B$ is the "direct sum" of matrices A and B.

$$L_{sr} = \begin{bmatrix} L_{af}(\theta) & L_{ak}(\theta) & L_{ah}(\theta) \\ L_{af}(\theta - 120^\circ) & L_{ak}(\theta - 120^\circ) & L_{ah}(\theta - 120^\circ) \\ L_{af}(\theta + 120^\circ) & L_{ak}(\theta + 120^\circ) & L_{ah}(\theta + 120^\circ) \end{bmatrix} , \tag{55}$$

$$L_{rr} = \left[\begin{array}{cc|c} L_{ff} & L_{fk} & \\ L'_{fk} & L_{kk} & 0 \\ \hline & 0 & L_{hh} \end{array}\right] = L'_{rr} . \tag{56}$$

The matrix L_{rr} is constant and symmetric nonnegative-definite, L_{ff} is the self-inductance of the field and L_{kk}, L_{hh} are square matrices whose orders are determined by the number of direct-axis and quadrature-axis damping bars, respectively. In addition,

$$L_{af} = M_f \cos\theta , \quad L_{ak} = M_k \cos\theta , \quad L_{ah} = -M_h \sin\theta \tag{57}$$

and

$$L_{aa} = L_s + M_0 \cos 2\theta , \quad L_{bc} = -M_s + M_0 \cos 2\theta \tag{58}$$

where M_k and M_h are 1-rowed matrices with non-negative elements and

$$M_f \geq 0; \quad L_s > M_0 \geq 0; \quad M_s > M_0 \geq 0 . \tag{59}$$

P_{k3}) Let

$$M_p(\theta) = \sqrt{\frac{2}{3}} \cdot \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \cos\theta & \cos(\theta - 120^\circ) & \cos(\theta + 120^\circ) \\ -\sin\theta & -\sin(\theta - 120^\circ) & -\sin(\theta + 120^\circ) \end{bmatrix} , \tag{60}$$

$$\Omega = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} = -\Omega' , \tag{61}$$

and

$$T_p(\theta) = M_p(\theta) \dotplus 1_\ell . \tag{62}$$

Then,

$$1_{3+\ell} = T'_p(\theta) T_p(\theta) , \tag{63}$$

$$C_0 = T'_p(0)(\Omega \dotplus 0_\ell) T_p(0) , \tag{64}$$

$$T(\theta, \mu) = T_0(\theta) = e^{-C_0\theta} \tag{65}$$

and

$$C_1 = T_p'(0) A_N T_p(0) \tag{66}$$

where

$$T_p'(0) = \sqrt{\frac{2}{3}} \cdot \begin{bmatrix} \frac{1}{\sqrt{2}} & 1 & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{1}{\sqrt{2}} & -\frac{1}{2} & -\frac{\sqrt{3}}{2} \end{bmatrix} + 1_\ell \tag{67}$$

and

$$A_N = \left[\begin{array}{ccc|cc} r_s L_0^{-1} & 0 & 0 & 0' & 0' \\ 0 & r_s(\ell_{f,k} L_d)^{-1} & 0 & -r_s\zeta_k' & 0' \\ 0 & 0 & r_s(\ell_{hh} L_q)^{-1} & 0' & -r_s\zeta_h' \\ \hline 0 & -R_{r_1}\zeta_k & -R_{r_2}\zeta_h & R_r\Delta & \end{array}\right] . \tag{68}$$

The entries in Eq. (68) are defined in terms of certain auxiliary quantities. Namely, let

$$L_0 = L_s - 2M_s \quad , \tag{69}$$

$$L_d = L_s + M_s + \frac{3}{2} M_0 \quad , \tag{70}$$

$$L_q = L_s + M_s - \frac{3}{2} M_0 \quad , \tag{71}$$

let

$$L_{f,k} = \left[\begin{array}{c|c} L_{ff} & L_{fk} \\ \hline L_{fk}' & L_{kk} \end{array}\right] \quad , \tag{72}$$

$$\eta_k' = \sqrt{\frac{3}{2}} [M_f \mid M_k] \quad , \tag{73}$$

$$\eta_h' = \sqrt{\frac{3}{2}} M_h \quad , \tag{74}$$

$$\ell_{f,k} = 1 - L_d^{-1}(\eta_k' L_{f,k}^{-1} \eta_k) \quad , \tag{75}$$

$$\ell_{hh} = 1 - L_q^{-1}(\eta_h' L_{hh}^{-1} \eta_h) \quad , \tag{76}$$

$$\Delta = L_d(L_d L_{f,k} - \eta_k \eta_k')^{-1} + L_q(L_q L_{hh} - \eta_h \eta_h')^{-1} \quad . \tag{77}$$

Then,

$$\zeta_k = \frac{1}{\ell_{f,k}} \cdot (L_{f,k} L_d)^{-1} \eta_k \quad , \tag{78}$$

$$\zeta_h = \frac{1}{\ell_{hh}} (L_{hh} L_q)^{-1} \eta_h \tag{79}$$

and

$$R_r = [R_{r_1} \mid R_{r_2}] \tag{80}$$

in which R_{r_1} and ζ_k' have the same number of columns and similarly for R_{r_2} and ζ_h'.

Joint Services Technical Advisory Committee
F44620-74-C-0056

D.C. Youla and J.J. Bongiorno, Jr.

National Science Foundation
ENG 74-13054

REFERENCES

1. E.W. Kimbark, "Power System Stability," (New York: John Wiley and Sons, Inc., 1956).

2. M.P. Erugin, "Linear Systems of Ordinary Differential Equations with Periodic and Quasi-Periodic Coefficients," (New York: Academic Press, 1966).

DESIGN OF INDUCTOR SYNCHRONOUS MOTORS

E. Levi, L. Birenbaum and Z. Zabar

Recently, new interest in the design of electrical machinery has been spurred by the development of solid state inverters, and by studies of vehicle propulsion using linear electric motors.[1] In this report, a new approach to the design of a special type of synchronous machine, the homopolar inductor motor, is described. This design method is especially suited for use with current-source-fed linear machines.

A. Construction and Control Aspects

Conventional synchronous machines are built with the field winding, excited by DC, located on the rotating part (rotor) while the AC armature winding is placed on the stationary part (stator). The inductor synchronous machine, on the other hand, has both the field and armature windings on the stator. Since the rotor has no winding, it can be made as a relatively sturdy and simple structure. Linear versions of this machine have been proposed for train propulsion. In this case, all of the windings may be placed on the undercarriage of the vehicle, and the track structure used as the counterpart of the rotor on the rotating machine. A sketch of an idealized linear inductor motor is shown in Figure 1.

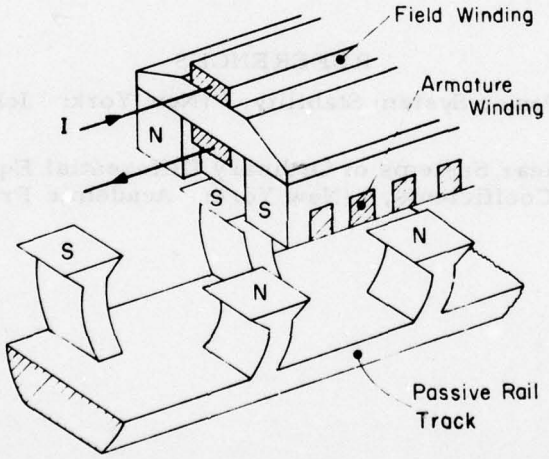


Fig. 1. Sketch of an idealized linear inductor motor.

The mechanism by which thrust is produced is the following: 3-phase currents are fed into the armature winding, and create a moving magnetic field. The frequency is so chosen that the linear speed of the field is precisely equal to the speed of motion, but oppositely directed. Hence the armature-induced field is at rest with respect to the track structure, which has poles induced by the field winding on the vehicle above it.

Only when the armature and field-produced magnetic configurations are at rest is thrust possible in a synchronous machine.

Speed control is achieved by varying the frequency of the 3-phase currents, supplied by a current-source inverter. Stable operation in this mode is obtained by use of a control system (Fig. 2) that accurately relates the inverter frequency to the vehicle speed, and in addition controls the phase angle between the track pole position and the armature current-induced magnetic field.[2, 3]

B. General Design Aspects

A list of symbols used in this, and the following sections, is given below.

A	armature surface area	I_f	DC current in the field winding
B'	rms fundamental normalized terminal voltage	K_a	rms armature surface current density
B_a	rms fundamental armature excited flux density $(\mu_o \tau / g\pi) K_a$	ℓ_p	pole face width
B_f	rms fundamental component of flux density in air gap due to field winding	N_f	turns/pole in the field winding
B_g	rms fundamental component of flux density in air gap	p	power per unit area of armature surface, P/A
B_{pf}	flux density at the pole face	t	thrust per unit area of armature surface
$B_{pf\ max}$	maximum allowable flux density at the pole face	v	velocity
$B_{t\ max}$	maximum allowable flux density in armature teeth	w_t	tooth width
C_d	X_{ad}/X_m, X_{ad} = direct axis armature reactance	X_f	leakage reactance of armature
C_q	X_{aq}/X_m, X_{aq} = quadrature axis armature reactance	X_m	magnetizing reactance
C_{lf}	ratio of maximum value of fundamental component of field-induced B to value under pole	α	$2\alpha = \pi(\ell_p/\tau)$ = angular span of pole face
d	distance from center of pole face	μ_o	permeability of free space $4\pi \times 10^{-7}$ H/m
g	effective air gap length	τ	pole pitch
		τ_s	slot pitch
		Φ'	angle between armature current and $\overline{B}_g$
		ψ	angle between direct axis and armature current maximum

Usually, an electrical machine is designed by a trial and error process. First, the geometry, dimensions and the magnetic and electric loading are chosen. From these are derived the terminal characteristics. These steps are then repeated until the design specifications are met. Here, instead, the starting point is the set of design specifications; these are then used to determine the magnetic and electric stresses; then the geometry, and finally the dimensions.

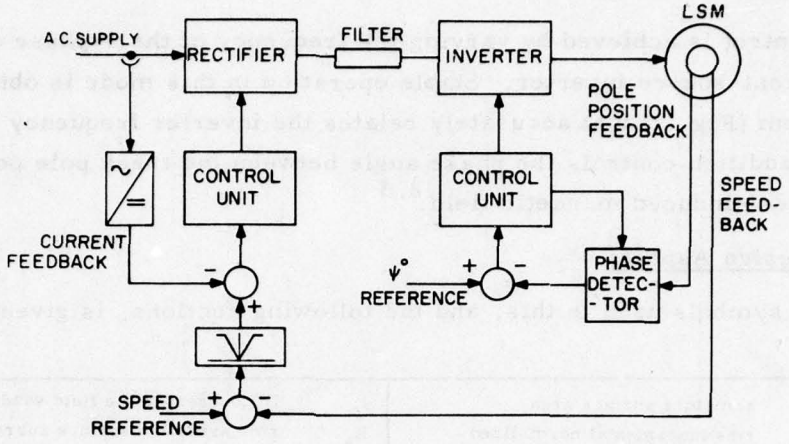


Fig. 2. Block diagram of a control system for the LSM.

In the inductor machine, an average component of B is present in addition to the useful alternating component. The flux density alternates around this average value, instead of between positive and negative peaks, as in a conventional synchronous machine. Hence the machine is prone to saturation of the magnetic flux path. The armature teeth are especially heavily stressed. It is important to avoid saturation of the salient pole structure, since the machine relies on variable reluctance effects for its operation.

The armature tooth saturation problem may be partially overcome by using a positive value for the angle ψ (see Fig. 3), between the direct axis and the maximum of the armature surface current density K_a. This causes the fundamental components of the armature and field-induced magnetic flux densities (B_a and B_f), or more precisely their mmf's, to subtract along the pole face, thus reducing tooth saturation in the section of the armature directly above the pole.

If the design is based on unity power factor, motor size is minimized for a given output power. This is done here.

C. Specific Design Aspects

As the starting point, we suppose that the maximum allowable flux density in the armature teeth ($B_{t,\,max}$) is specified. This, in turn, determines the maximum flux density at the pole face ($B_{pf,\,max}$):

$$B_{pf,\,max} = \frac{w_t}{\tau_s} B_{t,\,max} \tag{1}$$

Here, τ_s is the slot pitch and w_t the tooth width.

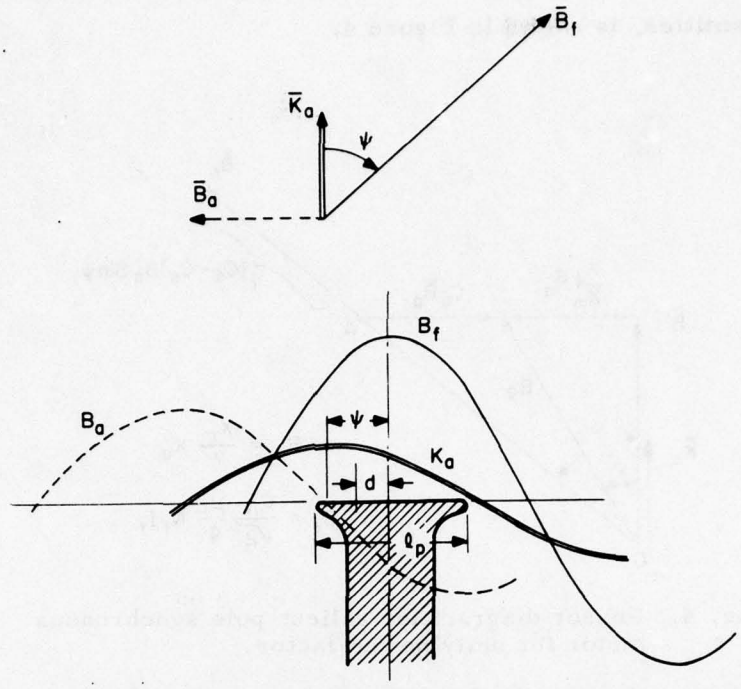


Fig. 3. Relation between fundamental components of armature and field produced magnetic fields for $\psi > 0$.

The flux density at any point above the pole face B_{pf} may be calculated using Ampere's law, and is found to be

$$B_{pf} = \frac{\mu_o}{g}\left[N_f I_f + \frac{\sqrt{2}\,\tau}{\pi} K_a \sin\left(\frac{\pi}{\tau} d - \psi\right)\right] \quad , \tag{2}$$

where the first term is related to the ampere-turns/pole, $N_f I_f$, and the second is related to the RMS armature surface current density K_a. (τ is the pole pitch, or half the distance between successive N poles; g is the effective air gap, i.e., the effective separation between the pole surface and the under surface of the armature core.) At the leading pole edge, where B_{pf} is greatest, $\frac{\pi}{\tau} d$ becomes equal to α, where $2\alpha \quad \pi \frac{\ell_p}{\tau}$ is the angular span of the pole face.

From Eq. (2), it is seen that a choice of $\psi > \alpha$ avoids the situation where the armature and field mmf's are additive. Also α should be reasonably small, since this decreases the permeance of the interpolar space, and hence reduces the transverse flux. As a result, the weights of the field yoke and of the pole structure are both reduced. An optimum value of α is about 50°.[4,5]

The Blondell 2-reactance theory[6] can now be used to describe the machine behavior. A phasor diagram, normalized to show the (RMS) fundamental components of the magnetic field quantities, is shown in Figure 4.

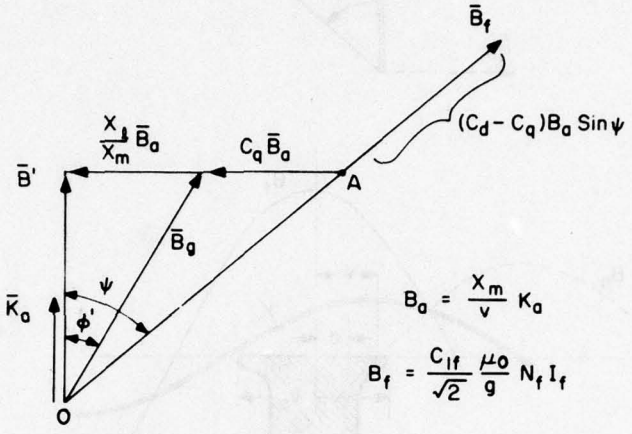


Fig. 4. Phasor diagram for salient-pole synchronous motor for unity power-factor.

From this phasor diagram the following identity can be derived:

$$\overline{OA} = \left(\frac{X_\ell}{X_m} + C_q\right) \frac{\mu_o \tau}{\pi g} \frac{1}{\sin\psi} K_a = B_f - (C_d - C_q) \frac{\mu_o \tau}{\pi g} K_a \sin\psi \tag{3}$$

and solving for K_a,

$$K_a = \frac{\pi g}{\mu_o \tau} \frac{B_f}{(C_d - C_q)\sin\psi + \left(\frac{X_\ell}{X_m} + C_q\right)\frac{1}{\sin\psi}} \tag{4}$$

A relation is thus established between K_a and B_f which are indices of the electric and magnetic loading, respectively.

If A is the area of the air gap, v the velocity of relative motion between the armature and the polar structure, the electromagnetic force developed per unit area of the air gap is:

$$t = \frac{P}{Av} = K_a \cdot B_g \cos\Phi' = K_a \cdot \overline{OA} \cos\psi \tag{5}$$

The terms C_d and C_q are proportional to the direct and quadrature saturated reactances, and their difference is present in Equation (4). The approximation is now

made that the first term in the denominator, $(C_d - C_q) \sin \psi$, is small in comparison with the second. This permits the following simplified expression for the thrust per unit area to be obtained when Eqs. (3) and (4) are substituted into Eq. (5):

$$t \cong \frac{\pi g}{\mu_o \tau} \frac{B_f^2}{\frac{X_\ell}{X_m} + C_q} \cdot \frac{\sin 2\psi}{2} \tag{6}$$

The force density is maximum for $\psi = 45^\circ$, a value which is not too different from the optimal value of α.

If ψ is set to a value equal to the angle α, no additive mmf will be present, even at the leading pole edge (Equation (2)). When α is close to its optimum value, this choice of ψ also maximizes the thrust-per-unit area, and hence, for a specified power output tends to lead to a machine of minimum size.

D. Conclusion

It has been shown that one may approach the design of an inductor motor by specifying an upper limit for the flux density in the armature teeth, and then proceeding, almost directly, to obtain a thrust per unit area, and hence a machine size.

Department of Transportation
DOT-FR-30030 and DOT-FR-64227

E. Levi, L. Birenbaum and Z. Zabar

REFERENCES

1. R.B. Powell, "Linear Induction Motor Electrical Performance Test," Final Report No. FRA-OR and D 76-265 by AiResearch Mfg. Co. for Federal Railroad Administration under Contract No. DOT-FR-40016 (June 1976). See also current Quarterly Reports, General Electric Co., under the SLEM program, for the Federal Railroad Administration, Contract No. DOT-FR-64147.
2. G.R. Slemon, S.B. Dewan, J.W.A. Wilson, "Synchronous Motor Drive with Current-Source Inverter," IEEE IA-10, No. 3, pp. 412-416 (May-June 1974).
3. Von Dieter Kollensperger, "Die Synchronmaschine als selbstgesteuerter Stromrichtermotor," Siemens-Zeitschrift 41, Heft 10 (1967).
4. E. Levi, "A Preliminary Evaluation of Electrical Propulsion by Means of Iron-Cored Synchronously Operating Linear Motors," Final Report for DOT-FRR, NTIS No. PB-258437 (January 1975).
5. E. Levi, "Preliminary Design Studies on Iron-Cored Synchronously Operating Linear Motors," Final Report for DOT-FRR, Report No. POLY EE/EP 76-005 (May 1976).
6. M. Kostenko and L. Piotrovsky, "Electrical Machines," (Moscow: Mir Publishers, 1974) 3rd Ed., Vol. 2.

II. SYSTEMS

A. COMMUNICATIONS

Program Director: L. Kurz

B. COMPUTERS AND COMPUTER-COMMUNICATION NETWORKS

Program Director: E.J. Smith

C. SAFETY, RELIABILITY AND SOFTWARE ENGINEERING

Program Director: M.L. Shooman

D. SYSTEMS, CONTROL AND NETWORKS

Program Director: D.C. Youla

E. DATA PROCESSING

Program Director: A. Papoulis

MAXIMUM A POSTERIORI ESTIMATOR FOR SUPPRESSION OF INTERCHANNEL INTERFERENCE IN FM RECEIVERS

F. A. Cassara and H. Schachter

A. Introduction

In a previous work[1] experimental evaluation of a novel FM detector capable of suppressing interchannel interference in FM receivers was presented. In this report it is theoretically demonstrated that the receiver structure experimentally studied represents the maximum a posteriori (MAP) estimator for the phase of the desired received FM signal when this signal is simultaneously corrupted by a co-channel interferer and additive white Gaussian noise.

B. Optimum Receiver

Consider a received FM signal with the interchannel interference and additive Gaussian noise which may be represented mathematically as

$$v(t) = s_1(t) + s_2(t) + n(t) , \quad 0 \leq t \leq T \tag{1}$$

where

$$s_1(t) = A_1 \cos \left(\omega_1 t + \int^t m_1(u)\, du \right)$$

$$s_2(t) = A_2 \cos \left(\omega_2 t + \int^t m_2(u)\, du \right)$$

$n(t)$ is white Gaussian noise with zero mean and normalized two sided power spectral density of 1 watt/Hz.

$m_1(t)$ is the desired signal modulation

$m_2(t)$ is the modulation on the interfering FM wave

A_1 and A_2 are assumed constant

$m_1(t)$ and $m_2(t)$ are assumed to be independent stationary Gaussian processes with zero mean and autocorrelation $R_{m_1}(\tau)$ and $R_{m_2}(\tau)$. We designate

$$x_1(t) = \int^t m_1(u)\, du \tag{2}$$

$$x_2(t) = \int^t m_2(u)\, du \tag{3}$$

This converts the FM waves $s_1(t)$ and $s_2(t)$ into angle modulated waves with $x_1(t)$ and $x_2(t)$ as modulating signals. It is seen that $x_1(t)$ and $x_2(t)$ are also mutually independent zero mean Gaussian random processes. Let the autocorrelation functions of $x_1(t)$ and $x_2(t)$ be denoted by $R_{x_1}(\tau)$ and $R_{x_2}(\tau)$.

We now derive a MAP receiver for the optimum reception of $m_1(t)$. Van Trees[2] has proved that a linear operation on the MAP estimate of a continuous random process is equal to the MAP estimate of the output of the linear operation performed on the random process. Therefore MAP estimation of $x_1(t)$ and subsequent differentiation of the estimate $\hat{x}_1(t)$ will give the MAP estimate $\hat{m}_1(t)$ of $m_1(t)$. We follow this procedure. We use abstract vector space methods throughout as developed by Schwartz.[3] We do not indicate the time-dependence of variables in the interest of brevity unless where explicitly necessary.

The controlling relation for MAP estimation of $x_1(t)$ is[3]

$$\frac{\partial p(x_1|v)}{\partial x_1} = 0 \tag{4}$$

where $p(x_1|v)$ is the conditional probability density function of x_1 given the received signal $v(t)$. Equation (4) can be written in the following equivalent form

$$\frac{\partial [\log p(x_1|v)]}{\partial x_1} = 0 \tag{5}$$

Using the relation

$$p(x_1|v) = p(v|x_1) \cdot p(x_1)/p(v)$$

and

$$\frac{\partial p(v)}{\partial x_1} = 0$$

we obtain

$$\frac{\partial \log p(v/x_1)}{\partial x_1} + \frac{\partial \log p(x_1)}{\partial x_1} = 0 \tag{6}$$

The first part of Eq. (6) takes the form

$$\frac{\partial}{\partial x_1} \log p(v|x_1) = \frac{1}{p(v|x_1)} \cdot \frac{\partial}{\partial x_1} \int_{-\infty}^{\infty} p(v|x_1, x_2)\, p(x_2)\, dx_2$$

$$= \frac{1}{p(v|x_1)} \frac{\partial}{\partial x_1} \int_{-\infty}^{\infty} p(x_2)\, k_1 \exp\left\{-\frac{1}{2}\left((v - s_1 - s_2), R_n^{-1}(v - s_1 - s_2)\right)\right\} dx_2 \tag{7}$$

where k_1 is a constant and (a, b) denotes inner product of a and b given by

$$(a, b) = \int_0^T a(t)\, b(t)\, dt.$$

The operator R_n^{-1} in Eq. (7) is defined by[3]

$$R_n^{-1}(v - s_1 - s_2) = \int_0^T R_n^{-1}(t, \mu) \left[v(\mu) - s_1(\mu) - s_2(\mu) \right] d\mu \tag{8}$$

or equivalently,

$$v - s_1 - s_2 = \int_0^T R_n(t, \mu) \left[R_n^{-1}(v(\mu) - s_1(\mu) - s_2(\mu) \right] d\mu$$

For white additive noise, $R_n(t, \mu) = \delta(t - \mu)$ so that

$$R_n^{-1}\left[v - s_1 - s_2 \right] = v - s_1 - s_2 \tag{9}$$

Substituting Eq. (9) into Eq. (7) and carrying out the indicated differentiation with respect to x_1 produces

$$\frac{\partial}{\partial x_1} \log p(v|x_1) = \frac{k_1}{p(v|x_1)} \int_{-\infty}^{\infty} p(x_2)\, p(v|x_1, x_2) \int_0^T (v - s_1) \frac{\partial s_1}{\partial x_1} dt\, dx_2$$

$$- \frac{k_1}{p(v|x_1)} \int_{-\infty}^{\infty} p(x_2)\, p(v|x_1, x_2) \int_0^T s_2 \frac{\partial s_1}{\partial x_1} dt\, dx_2 \tag{10}$$

The first expression in Eq. (10) is easily simplified by integrating with respect to x_2. In the second expression we note that, if x_1 and x_2 are statistically independent, we have

$$\frac{p(x_2)\, p(v|x_1, x_2)}{p(v|x_1)} = p(x_2|v, x_1) \tag{11}$$

With this, Eq. (10) is reduced to

$$\frac{\partial}{\partial x_1} \log p(v|x_1) = k_1 \int_0^T (v - s_1) \frac{\partial s_1}{\partial x_1} dt - k_1 \int_0^T \bar{s}_2 \frac{\partial s_1}{\partial x_1} dt \tag{12}$$

where

$$\bar{s}_2(t) = \int_{-\infty}^{\infty} p(x_2|v, x_1)\, s_2\, dx_2 \tag{13}$$

is recognized as the conditional mean and hence the minimum mean square estimate of $s_2(t)$.

Now following the second term in Eq. (6)

$$\frac{\partial}{\partial x_1} \log p(x_1) = \frac{\partial}{\partial x_1} \log \left[k_2 \exp \left[-\frac{1}{2} \left(x_1, R_{x_1}^{-1} x_1 \right) \right] \right] = -\frac{1}{2} \frac{\partial}{\partial x_1} \left(x_1, R_{x_1}^{-1} x_1 \right) = -R_{x_1}^{-1} x_1 \tag{14}$$

Again making use of Eq. (8) to simplify Eq. (14) and adding the result to Eq. (12) as per Eq. (6) we have after an inverse transformation

$$x_1(t) = k_1 \int_o^T R_{x_1}(t, u) \frac{\partial s_1}{\partial x_1} (v - s_1 - \bar{s}_2) \, du \tag{15}$$

Solutions of Eq. (15) for $x_1(t)$ are the MAP estimates, we denote these by $\hat{x}_1(t)$. Further, in Eq. (15) it is noted that the integral on R.H.S. is a convolution integral in which the filter $R_{x_1}(t, u)$ is low-pass, whereas the term $s_1 \, \partial s_1 / \partial x_1$ involves double the carrier frequency terms and hence would integrate to zero. Equation (15) therefore simplifies to

$$\hat{x}_1(t) = k_1 \int_o^T R_{x_1}(t, u) \frac{\partial \hat{s}_1}{\partial x_1} \cdot (v - \bar{s}_2) \, du \tag{16}$$

Equation (16) clearly shows a phase lock loop structure with $\bar{s}_2(t)$ as an external input as shown in Figure 1.

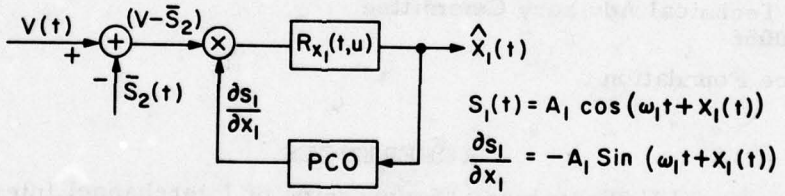


Fig. 1 Optimum receiver for $x_1(t)$.

Following all steps from Eq. (7) through Eq. (16) but estimating $x_2(t)$ instead of $x_1(t)$ we obtain

$$\hat{x}_2(t) = k_2 \int_o^T R_{x_2}(t, u) \frac{\partial \hat{s}_2}{\partial x_2} \cdot (v - \bar{s}_1) \, du \tag{17}$$

This is a receiver identical to Fig. 1 except that $\bar{s}_2$ is replaced by $\bar{s}_1$ and $R_{x_1}(t, \mu)$ is replaced by $R_{x_2}(t, u)$. If we assume $\bar{s}_1$ and $\bar{s}_2$ are suitable estimates of $\hat{s}_1$ and $\hat{s}_2$, respectively, i.e., the m.m.s.e. are approximately equal to the MAP estimates,

then the receiver of Fig. 1 and its equivalent for obtaining $\hat{x}_2(t)$ can be coupled to obtain a comprehensive receiver with no unknown inputs. Recalling that in the case of FM, the modulation functions are $m_1(t) = \dot{x}_1(t)$ and $m_2(t) = \dot{x}_2(t)$ and replacing the phase controlled voltage controlled oscillator (VCO) of Fig. 1 by a frequency controlled VCO, such a cross coupled FM receiver appears as in Figure 2. It must be emphasized that only when $\hat{s}_1(t) = \bar{s}_1(t)$ and $\hat{s}_2(t) = \bar{s}_2(t)$ will this receiver be optimum.

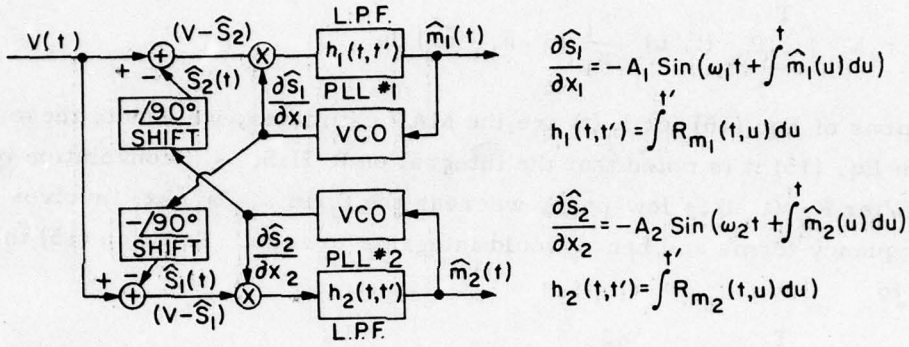


Fig. 2. Novel FM detector for suppression of interchannel interference.

Figure 2 describes the physical receiver structure experimentally evaluated during the previous research period and discussed in Reference 1.

Joint Services Technical Advisory Committee
F44620-74-C-0056 F. A. Cassara

National Science Foundation
ENG 76-24468

REFERENCES

1. F. A. Cassara and H. Schachter, "Suppression of Interchannel Interference in FM Receivers," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76, pp. 300-304 (1976).

2. H. L. Van Trees, "Detection, Estimation and Modulation Theory," (New York: John Wiley and Sons, Inc., 1968).

3. Mischa Schwartz, "Abstract Vector Spaces Applied to Problems in Detection and Estimation Theory," Trans. IEEE on Information Theory, pp. 327-336 (July 1966).

A NEW APPROACH TO THE EDGEWORTH SERIES APPROXIMATION WITH APPLICATION TO THE BINOMIAL DISTRIBUTION

C. Y. Chang and L. Kurz

In evaluating the performance of some communication, radar and sonar systems, as well as in reliability and quality control work, frequently it is necessary to find accurate expressions for the binomial distribution. In the past, Edgeworth series approximations[1] have been used with success but this approach fails when the probability of success $p \leq .2$ for the four-term series calculated so far. In this report, firstly, the Edgeworth series is extended to include nine terms, which yields excellent accuracy for $p \geq .1$ and $m \geq 36$; secondly, a technique of shift-accuracy-interval (TSAI) is introduced which guarantees high accuracy even for a four-term Edgeworth series (10^{-7} percent error is easily achievable). The technique used to generate $P_{m,n}$ polynomials for the binomial distribution is sufficiently general to indicate applicability to other distributions.

A. The Edgeworth Series

Consider the sum of n i.i.d. r.v.

$$\xi = \sum_{\nu=1}^{n} \xi_\nu$$

with means, m_1, and variance σ_1. Introducing the standardized r.v. $x = \xi - m/\sigma$, where $m = nm_1$ and $\sigma = \sigma_1\sqrt{n}$, we have an approximation to the p.d.f. of x

$$f(x) = c_o\phi(x) + \frac{c_1}{1!}\phi'(x) + \frac{c_2}{2!}\phi''(x) - \frac{c_3}{3!}\phi'''(x) + \cdots \tag{1}$$

where

$$\phi^i(x) = \frac{(-1)^i}{\sqrt{2\pi}} e^{-x^2/2} M_i(x), \quad i = 0, 1, 2, \cdots, \phi^o(x) = \phi(x) \tag{2}$$

and $M_i(x)$ are the Hermite polynomials

$$H_i(x) = \sum_j (-\tfrac{1}{2})^j x^{i-2j} \frac{i!}{j!(i-2j)!} = x^i + h_{i,(i-2)}x^{i-2} + h_{i,(i-4)}x^{i-4} + \cdots \tag{3}$$

where the sum is taken from $j = 0$ to the longest value for which $i - 2j \geq 0$. Equation (1) is then

$$f(x) = [c_o + \frac{c_1}{1!}M_1(x) + \frac{c_2}{2!}M_2(x) + \frac{c_3}{3!}M_3(x) + \cdots]\phi(x) \tag{4}$$

with

$$c_i = \int_{-\infty}^{\infty} M_i(x) f(x)\, dx \tag{5}$$

Because f(x) is of zero mean and unit variance, while its higher order moments are μ_ν/σ^ν, c_i may be expressed by

$$c_i = \frac{\mu_i}{\sigma_i} + h_{i,(i-2)} \frac{\mu_{i-2}}{\sigma^{i-2}} + h_{i,(i-4)} \frac{\mu_{i-4}}{\sigma^{i-4}} + \cdots \quad \nu = 1, 2, \cdots \tag{6}$$

Denoting by η_ν the semi-invariants of $(\xi - m)$ and by η'_ν the semi-invariants of $\xi_1 - m_1$ and introducing

$$\lambda_\nu = \frac{\eta_\nu}{\sigma^\nu} \quad \text{and} \quad \lambda'_\nu = \frac{\eta'_\nu}{\sigma_1^\nu},$$

we obtain

$$f(x) = \phi(x) + \sum_{\nu=1}^{\infty} (-1) \frac{b_{\nu,\nu+2}\, \phi^{(\nu+2)}(x) + \cdots + b_{\nu,3\nu}\, \phi^{3\nu}(x)}{n^{\nu/2}} \tag{7}$$

where $b_{\nu,\nu+2h}$ is a polynomial in $\lambda'_3, \cdots, \lambda'_{\nu-h+3}$ independent of n. This is the Edgeworth series.

Applying the Edgeworth series to the binomial distribution, one should keep in mind that it is a discrete distribution while f(x) represents a p.d.f., a continuous distribution. The approximation is meaningful in terms of c.d.f., or the integral replaces the same as an approximant. Consider the binomial distribution with the f.f.

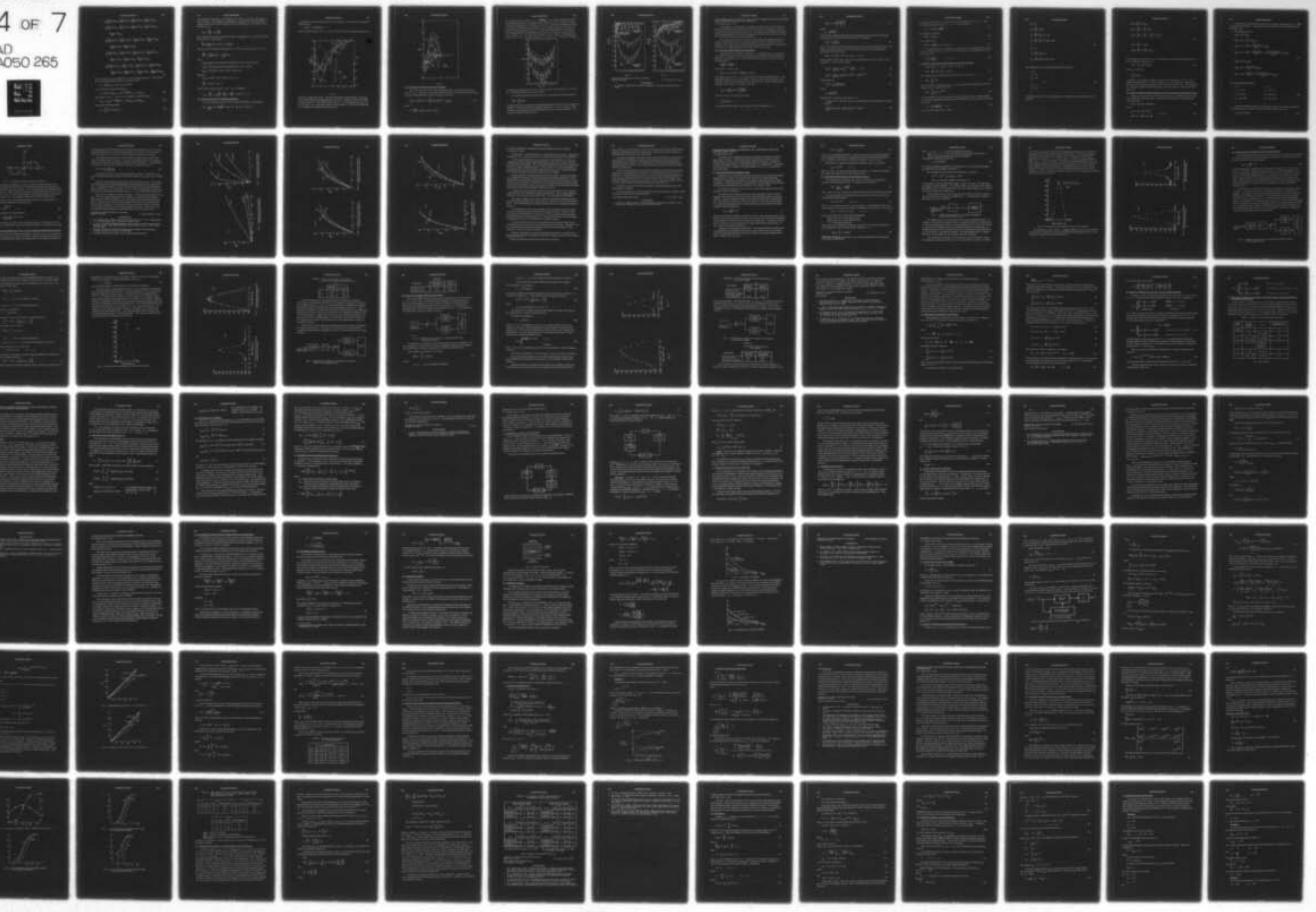
$$B_p(k) = \binom{m}{k} p^k q^{m-k} \tag{8}$$

The first nine terms of the Edgeworth series in terms of powers of $1/\sigma$ may be written as

$$\begin{aligned} f(x) = \phi(x) &- \frac{1}{\sigma}\left[\frac{P_{1,3}}{3!}\, \phi^{(3)}(x)\right] \\ &+ \frac{1}{\sigma^2}\left[\frac{P_{2,4}}{4!}\, \phi^{(4)}(x) + \frac{P_{2,6}}{6!}\, \phi^{(6)}(x)\right] \\ &- \frac{1}{\sigma^3}\left[\frac{P_{3,5}}{5!}\, \phi^{(5)}(x) + \frac{P_{3,7}}{7!}\, \phi^{(7)}(x) + \frac{P_{3,9}}{9!}\, \phi^{(9)}(x)\right] \end{aligned}$$

AD-A050 265 POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 20/3
PROGRESS REPORT NO. 42, 15 SEPTEMBER 1976 - 14 SEPTEMBER 1977, --ETC(U)
NOV 77 A A OLINER F44620-74-C-0056
UNCLASSIFIED POLY-MRI-452.42-77 NL

4 OF 7
AD
A050 265



$$+\frac{1}{\sigma^4}\left[\frac{P_{4,6}}{6!}\phi^{(6)}(x)+\frac{P_{4,8}}{8!}\phi^{(8)}(x)+\frac{P_{4,10}}{10!}\phi^{(10)}(x)+\frac{P_{4,12}}{12!}\phi^{(12)}(x)\right]$$

$$-\frac{1}{\sigma^5}\left[\frac{P_{5,7}}{7!}\phi^{(7)}(x)+\frac{P_{5,9}}{9!}\phi^{(9)}(x)+\frac{P_{5,11}}{11!}\phi^{(11)}(x)+\frac{P_{5,13}}{13!}\phi^{(13)}(x)+\right.$$

$$\left.\frac{P_{5,15}}{15!}\phi^{(15)}(x)\right]$$

$$+\frac{1}{\sigma^6}\left[\frac{P_{6,8}}{8!}\phi^{(8)}(x)+\frac{P_{6,10}}{10!}\phi^{(10)}(x)+\frac{P_{6,12}}{12!}\phi^{(12)}(x)+\frac{P_{6,14}}{14!}\phi^{(14)}(x)+\right.$$

$$\left.\frac{P_{6,16}}{16!}\phi^{(16)}(x)+\frac{P_{6,18}}{18!}\phi^{(18)}(x)\right]$$

$$-\frac{1}{\sigma^7}\left[\frac{P_{7,9}}{9!}\phi^{(9)}(x)+\frac{P_{7,11}}{11!}\phi^{(11)}(x)+\frac{P_{7,13}}{13!}\phi^{(13)}(x)+\frac{P_{7,15}}{15!}\phi^{(15)}(x)+\right.$$

$$\left.\frac{P_{7,17}}{17!}\phi^{(17)}(x)+\frac{P_{7,19}}{19!}\phi^{(19)}(x)+\frac{P_{7,21}}{2!}\phi^{(21)}(x)\right]$$

$$+\frac{1}{\sigma^8}\left[\frac{P_{8,10}}{10!}\phi^{(10)}(x)+\frac{P_{8,12}}{12!}\phi^{(12)}(x)+\frac{P_{8,14}}{14!}\phi^{(14)}(x)+\frac{P_{8,16}}{16!}\phi^{(16)}(x)+\right.$$

$$\left.\frac{P_{8,18}}{18!}\phi^{(18)}(x)+\frac{P_{8,20}}{20!}\phi^{(20)}(x)+\frac{P_{8,22}}{22!}\phi^{(22)}(x)+\frac{P_{8,24}}{24!}\phi^{(24)}(x)\right] \tag{9}$$

The coefficients $P_{m,n}$ are independent of n and are polynomials in p. The moments may be evaluated using the moment-generating function

$$m_i = G_x^{(i)}(s)\Big|_{s=0}, \text{ where } G_x(s) = E[e^{sk}]$$

It is convenient to write the moments as

$$m_i = A_i p^i + B_i(s)\Big|_{s=0} = A_i p^i + \frac{(i-1)i}{2} m_{i-1} + C_i(s)\Big|_{s=0} \tag{10}$$

where $A_i(n) = n(n-1)\cdots[n-(i-1)]$. Finally, in a recursive form

$$m_{i+1} = A_{i+1} p^{i+1} + \frac{i(i+1)}{2} m_i - i B_i(s)\Big|_{s=0} + C_i'(s)\Big|_{s=0} \tag{11}$$

The central moments are then

$$\mu_k = \sum_{r=0}^{k} \binom{k}{r} (-1)^r m_1^r m_{k-r} \tag{12}$$

which permits us to find $P_{m,n}$ from Equation (6). All $P_{m,n}$ from $P_{1,3}$ through $P_{8,24}$ were calculated. The tabulation is not included in this report because of its length; instead, the procedure will be demonstrated by calculating $P_{2,6}$. From Eqs. (4), (6) and (9), $P_{2,6}$ is

$$P_{2,6} = [\frac{\mu_6}{\sigma^4}] - h_{6,4}[\frac{\mu_4}{\sigma^2}]$$

where $[\cdot]$ that part of the bracketed term which is independent of n. From Eq. (12) and the fact that $\sigma^2 = npq$, we have

$$\frac{\mu_6}{\sigma^4} = \frac{1}{n^2p^2q^2}(m_6 - 6m_1m_5 + 15m_1^2m_4) .$$

The m_r may be calculated from Equation (11). They are of the form $m_r = n[\cdots]$, or

$$[\frac{\mu_6}{\sigma^4}] = \frac{1}{n^2p^2q^2}[m_6]_2 - 6\frac{1}{npq^2}[m_5]_1$$

where

$[m_6]_2$ denotes the contribution of polynomial in p from m_6 of the form

$[m_6]_2 = n^2$ (polynomial in p independent of n). Thus, we have

$$[m_6]_2 = n^2(274p^4 - 750p^3 + 715p^2 - 270p + 31]$$

Similarly

$$[m_5]_1 = n(24p^4 - 60p^3 + 50p^2 - 15p + 1), \text{ or}$$

$$[\frac{\mu_6}{\sigma^4}] = 5(26p^2 - 26p + 5)$$

By the same procedure $[\mu_4/\sigma^2] = 6p^2 - 6p + 1$, obtaining

$$P_{2,6} = [\frac{\mu_6}{\sigma^4}] - h_{6,4}[\frac{\mu_4}{\sigma^2}] = [\frac{\mu_6}{\sigma^4}] - 15[\frac{\mu_4}{\sigma^2}] = 10(4p^2 - 4p + 1)$$

B. Testing the Accuracy of the Edgeworth Series

To use the Edgeworth series for the binomial distribution, replace $\phi(x)$ by

$$\phi(k) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp - \frac{(k - \overline{m})^2}{2\sigma^2} \quad \text{where} \quad \overline{m} = mp \text{ and } \sigma^2 = mp(1 - p)$$

Let $B_a(k)$ be the series approximation of the actual $B_p(k)$. Introducing as a measure of approximation

$$|D_p(k)| = |B_a(k)/B_p(k) - 1|$$

Figure 1 shows the error for p = .1 and m = 36 with ℓ in (ℓ) indicating the number of

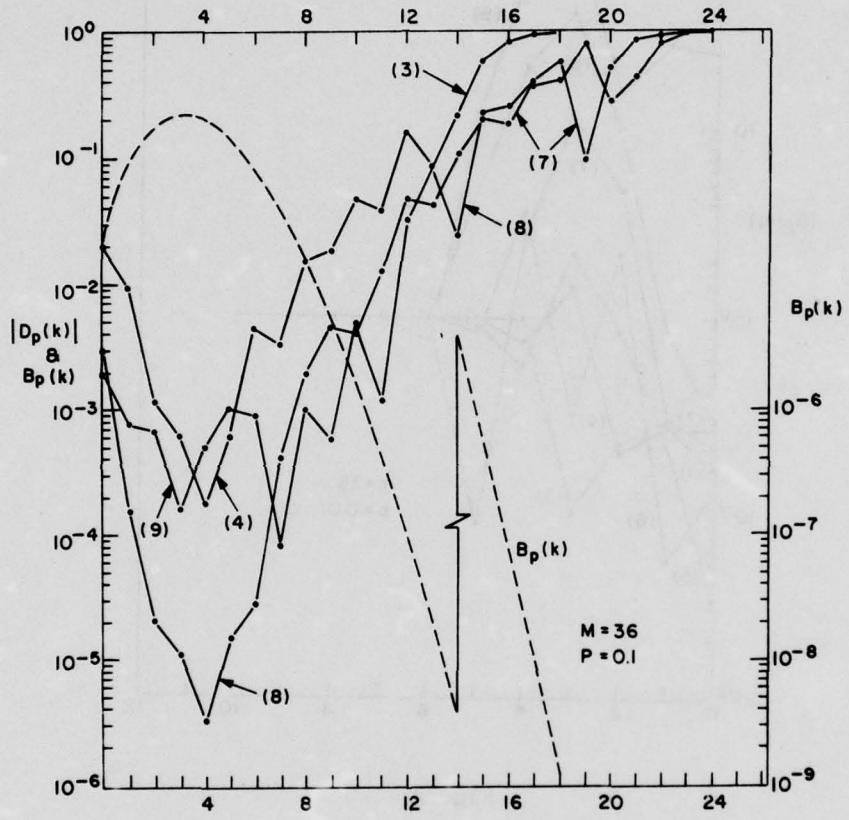


Fig. 1.

terms in the Edgeworth series. The figure demonstrates that the higher order Edgeworth series improves the approximation in two ways: improved accuracy and larger trustable range of k. This method breaks down if $\sigma < 1$ as demonstrated in Fig. 2 for p = .01 and m = 36. In the latter case the series diverges. To remove this difficulty, a new technique is presented in the next section.

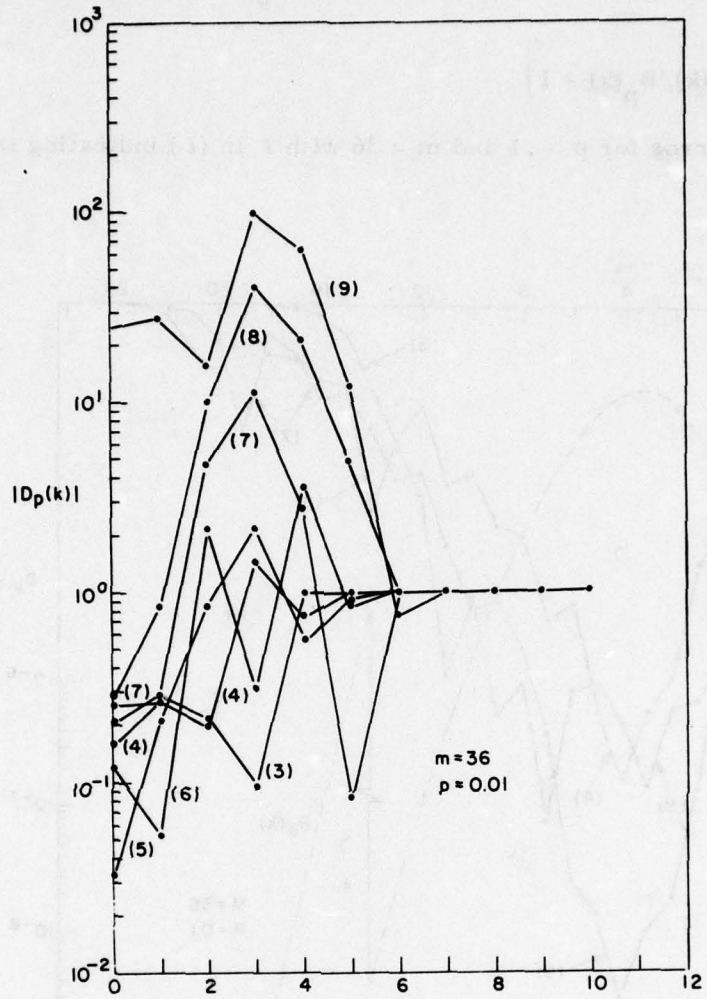


Fig. 2.

C. A Technique of Shift-Accuracy-Interval (TSAI)

Even if one uses nine terms of the Edgeworth series, the approximation breaks down for $\sigma \leq 1$. A powerful, yet simple, technique will now be introduced to ensure high accuracy of approximation. Let $p_1 = ap$ and $q_1 = q/b$, then

$$B_p(k) = \binom{m}{k} p^k q^{m-k} = \frac{b^{m-k}}{a^k} \binom{m}{k} p_1^k q^{m-k} = C\, B_{p_1}(k) \tag{13}$$

where

$$C = \frac{b^{m-k}}{a^k} = \exp[(m-k)\ln b - k \ln a]$$

By a suitable choice of a and b, i.e., $\sigma = \sqrt{mp_1q_1} > 1$, we reduce the problem to finding an accurate expression for $B_{p_1}(k)$ using the Edgeworth series, while the constant, C, may be calculated with a high degree of accuracy. Thus, by an appropriate change in p_1 and q_1, we obtain an accurate expression for $B_p(k)$ for every k. From Fig. 1 it is obvious that it suffices to use a four- or eight-term Edgeworth series for high accuracy representation in the neighborhood of the mean. In general, a good rule is to use p_1 so that $\sigma_1 > 0$, then TSAI will yield highly accurate results. The results are shown in Fig. 3, where TSAI was applied to p = .1 and m = 36. In the figure, $|D_p(k)|$

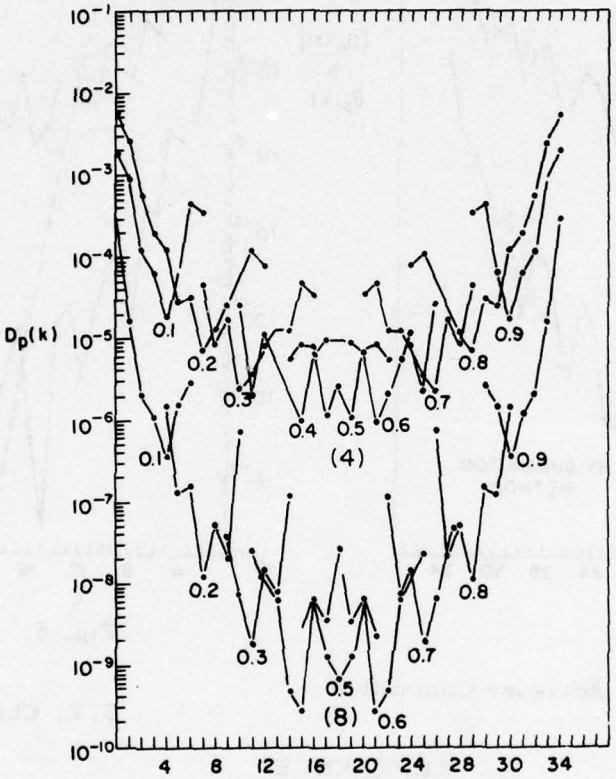


Fig. 3.

was plotted for different choices of p_1. Similar results were obtained for p = .01 and m = 36, therefore the graphs are not repeated.

If one needs high accuracy expressions for the cumulative distribution function

$$\theta_p(k) = \sum_{k=o}^{m} B_p(k)$$

we select the values of $B_p(k)$ from the appropriate high accuracy intervals for $B_p(k)$ and add them. The accuracy of this method was tested for p = .1 and .01, and m = 36 by selecting five values from the first shift interval and four from other intervals. The results are shown in Figures 4 and 5.

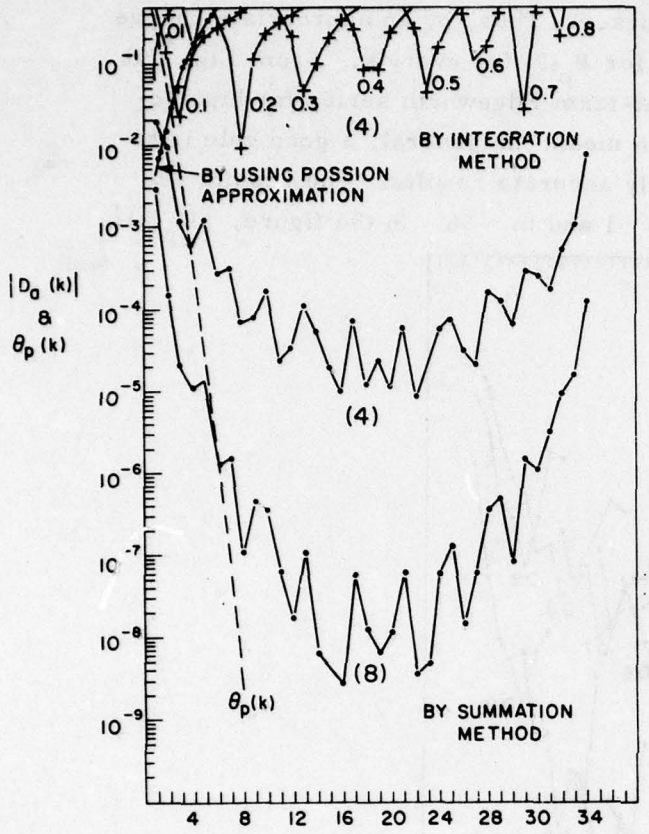


Fig. 4.

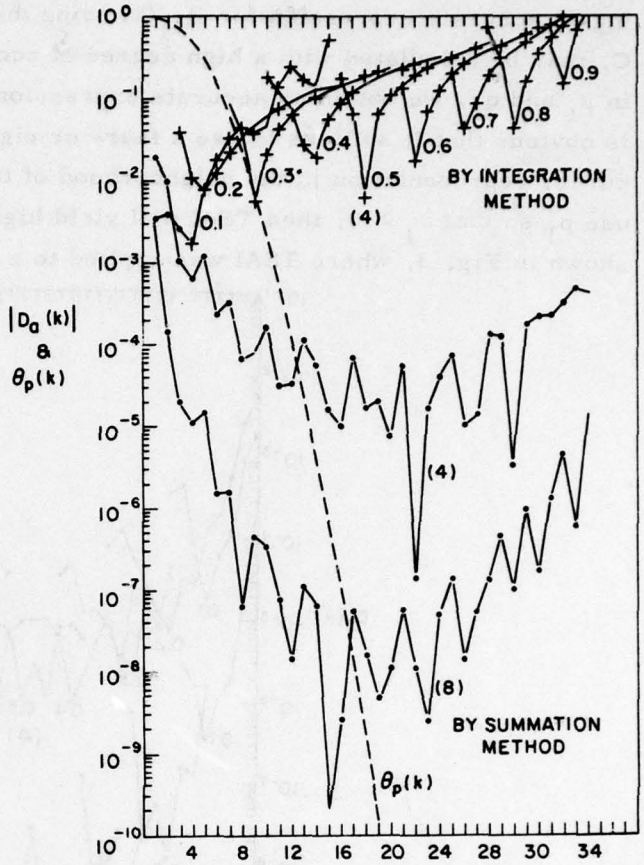


Fig. 5.

Joint Services Technical Advisory Committee
F44620-74-C-0056

C.Y. Chang and L. Kurz

REFERENCE

1. H. Cramer, "Mathematical Methods of Statistics," (Princeton University Press, 1946).

AUTOCORRELATION FUNCTION AT THE OUTPUT OF MEMORYLESS NONLINEARITIES USING A DIFFERENTIAL EQUATION APPROACH

R. Chassaing and L. Kurz

In this report, a new approach is presented for finding near-optimum detectors for noise distributions represented by a mixture of Gaussian and impulsive noise and assuming the noise processes are Markovian. The approach concentrates on minimizing the output noise power by solving the appropriate Fokker-Planck equation in conjunction with the Fletcher-Powell search procedure.

The autocorrelation function obtained using the differential equation approach is simpler to evaluate than using the standard transform method. Furthermore, the approach of this report can be extended naturally to obtain expressions for multi-dimensional p.d.f.'s, spectral density functions and signal-to-noise ratio.

A. The Fokker-Planck Equation and Its Solution

Using the notation suggested in Reference 1. In finding the p.d.f.'s associated with a Markov process, one may use the solution of the Fokker-Planck equation as an expression of the desired p.d.f. The Fokker-Planck equation is a second order differential equation

$$\frac{\partial W(x,t)}{\partial t} + \frac{\partial G(x,t)}{\partial x} = 0 \tag{1}$$

where $W(x,t)$ is a p.d.f. and

$$G(x,t) = K_1(x,t)\, W(x,t) - \frac{1}{2}\frac{\partial}{\partial x} K_2(x,t)\, W(x,t)$$

The coefficients $K_s(t,x)$, $s = 1, 2$, are the intensity coefficients defined in Ref. 1, and depending on their choices various Markov noise processes may be desired. If the intensity coefficients do not depend on time, then the p.d.f., $W(x,t)$, approaches its stationary value, $W_{st}(x)$. It is easy to show that

$$W_{st}(x) = \frac{A}{K_2(x)} \exp\left\{2\int_{x_1}^{x} \frac{K_1(y)}{K_2(y)}dy\right\} \tag{2}$$

where the choice of A and x_1 must satisfy

$$\int_{-\infty}^{\infty} W_{st}(x)\,dx = 1 .$$

For $W_{st}(x)$ Gaussian, $K_1(x) = ax + b$, $K_2(x) = d$ and the resulting p.d.f. is

$$W_{st}(x) = K \exp\left[\frac{(x+\frac{b}{a})^2}{\frac{d}{a}}\right] \tag{3}$$

where

$$K = 1/\sqrt{2\pi d/a} \quad .$$

To describe the amplitude distribution of the impulsive noise component, the $W_{st}(x)$, corresponding to the generalized Cauchy noise, is introduced

$$W_{st}(x) = \frac{K}{(x^2+h)^m} \tag{4}$$

For $m = 1$ we have the Cauchy distribution and for $m \to \infty$ the Gaussian distribution. The choice of m between these two limits represents a measure of the impulsiveness. Solving Eqs. (2) and (4) yields

$$K_1(x) = ax, \quad K_2(x) = c(x^2 + h) \quad \text{and} \quad c = a/(1-m).$$

A more general form of W_{st} of Eq. (4) may be obtained by the choice of $K_1(x) = ax^{n-1}$ and $K_2(x) = c$. For n even, we obtain

$$W_{st}(x) = \frac{K}{c(x^n+h)} \exp\{2\int_h^{x^n+h} \frac{a}{cn}\frac{du}{u}\} \quad u = x^n + h \tag{5}$$

$$W_{st}(x) = \frac{K}{c(x^n+h)} \exp\{\ell n(x^n+h)^{2a/cn} - \ell n(h)^{2a/cn}\} \tag{6}$$

which reduces to

$$W_{st}(x) = \frac{B}{(x^n+h)^{1-2a/cn}} \tag{7}$$

where

$$B = \frac{Kh^{-2a/cn}}{c} \quad .$$

Note that Eq. (7) reduces to Eq. (4) for $n = 2$.

Letting $W(x,t) = X(x)\,T(t)$ in Eq. (1), dividing by $W(x,t)$ and introducing a constant $\lambda = T(t)/T(t)$,

$$\frac{1}{2}\frac{\partial^2}{\partial x^2}[K_2(x)\,X(x)] - \frac{\partial}{\partial x}[K_1(x)\,X(x)] + \lambda\,X(x) = 0 \tag{8}$$

The solution of Eq. (8) consists of a sequence of eigenfunctions $X_0, X_1, X_2, \cdots$, with corresponding eigenvalues $\lambda_0, \lambda_1, \lambda_2, \cdots$. It can be shown that

$$(\lambda_m - \lambda_n) \int X_m X_n \frac{dx}{W_{st}(x)} = 0 \tag{9}$$

if the boundary conditions

$$G[X] = G[W_{st}] = 0 \tag{10}$$

are satisfied, where

$$G[U] = -\frac{1}{2}\frac{\partial}{\partial x}[K_2(x)\,U] + K_1(x)\,U \tag{11}$$

The orthogonality of the eigenfunctions with respect to the weight function $1/W_{st}(x)$ follows from Eq. (9) for $\lambda_m \neq \lambda_n$, $(m \neq n)$. Furthermore, each eigenfunction is normalized such that

$$\int_{-\infty}^{\infty} X_n^2 \frac{dx}{W_{st}(x)} = 1 \qquad n = 0, 1, 2, \cdots \tag{12}$$

B. Solution of the Fokker-Planck Equation for the Gaussian Case

Letting $K_1(x) = ax$, $K_2(x) = -2a\sigma^2 = d$ in Eq. (8)

$$X''(x) - \frac{2a}{d}\,x\,X'(x) + \frac{2}{d}(\lambda - a)\,X(x) = 0 \tag{13}$$

The solution of Eq. (13) about the ordinary point $x = 0$ is of the form

$$X(x) = \sum_{n=o}^{\infty} A_n x^n \tag{14}$$

with two arbitrary constants A_0 and A_1. A_n can be found by substituting Eq. (14) into Equation (13). A solution of the form

$$X(x) = \sum_{n=o}^{\infty} B_n x^n e^{-x^2/2\sigma^2} \tag{15}$$

can be assumed, or Eq. (14) can be used to generate Equation (15). Expanding Eqs. (14) and (15), equating like coefficients, the coefficients B_n can be found in terms of the A_n, where

$$A_n = -\frac{2}{d}\left(\frac{a - na + \lambda}{n(n-1)}\right) \qquad n \geq 2 \tag{16}$$

The B_n can be written with $d = -2a\sigma^2$

$$
\begin{aligned}
B_0 &= A_0 \\
B_2 &= \frac{A_0}{2!}\left(-\frac{2}{d}\right)\{\lambda\} \\
B_4 &= \frac{A_0}{4!}\left(-\frac{2}{d}\right)^2\{(\lambda+2a)\} \\
B_6 &= \frac{A_0}{6!}\left(-\frac{2}{d}\right)^3\{(\lambda+2a)(\lambda+4a)\} \\
&\vdots \\
B_1 &= A_1 \\
B_3 &= \frac{A_1}{3!}\left(-\frac{2}{d}\right)\{\lambda+a\} \\
B_5 &= \frac{A_1}{5!}\left(-\frac{2}{d}\right)^2\{(\lambda+a)(\lambda+3a)\} \\
&\vdots
\end{aligned}
\tag{17}
$$

The eigenvalues can be found by inspection from the B_n's,

$$
\begin{aligned}
\lambda_0 &= 0 \\
\lambda_2 &= -2a \\
\lambda_4 &= -4a \\
&\vdots \qquad\qquad a < 0 \\
\lambda_1 &= -a \\
\lambda_3 &= -3a \\
&\vdots
\end{aligned}
\tag{18}
$$

Using Eq. (17) in Eq. (15), the eigenfunctions corresponding to the eigenvalues of Eq. (18) are,

$$X_0(x) = \frac{A_0}{K}\{1\}\, W_{st}(x)$$

$$X_2(x) = \frac{A_0}{K}\{1 - \frac{x^2}{\sigma^2}\}\, W_{st}(x)$$

$$X_4(x) = \frac{A_0}{K}\{1 - \frac{2}{\sigma^2}x^2 + \frac{1}{3\sigma^4}x^4\}\, W_{st}(x)$$

$$\vdots \qquad (19)$$

$$X_1(x) = \frac{A_1}{K}\{x\}\, W_{st}(x)$$

$$X_3(x) = \frac{A_1}{K}\{x - \frac{x^3}{3\sigma^2}\}\, W_{st}(x)$$

$$\vdots$$

where $W_{st}(x)$ is the Gaussian p.d.f. Rearranging Eq. (19) yields the eigenfunctions in terms of Hermite polynomials, or

$$X_n(x) = \sum_n C_n H_n(x)\, W_{st}(x) \qquad (20)$$

where C_n *is such that*

$$\int_{-\infty}^{\infty} W_{st}(x)\, dx = 1$$

is satisfied. The boundary conditions, Eq. (16), were satisfied for $x = \pm\infty$. If the boundary condition $G[X_n] = 0$, n odd, is applied for $x \geq 0$, the arbitrary constant A_1 is zero, and the solution is expressible in terms of even polynomials only. The differential Eq. (13) has no singularity in the finite plane, hence its solution is valid for all finite values of x.

C. Solution of the Fokker-Planck Equation for the Generalized Cauchy Density Function

Using a similar procedure as for the Gaussian case, substituting $K_1(x)$ and $K_2(x)$ which satisfy Eq. (4) into Eq. (16), the Fokker-Planck equation for the generalized Cauchy becomes

$$X''(x) + p(x)\, X'(x) + q(x)\, X(x) = 0 \qquad (21)$$

where

$$p(x) = 2x(2 - \frac{a}{c})/(x^2 + \frac{d}{c})$$

$$q(x) = 2(1 - \frac{a+\lambda}{c})/(x^2 + \frac{d}{c}) \qquad \frac{d}{c} = h > 0 \qquad (22)$$

This second order differential equation has two regular singular points at $x = \pm j\sqrt{h}$ in the finite complex plane. The solution of Eq. (22) yields the following eigenfunctions and eigenvalues.

The eigenfunctions are

$$X_0(x) = A\{1\}\, W_{st}(x)$$

$$X_2(x) = A\left\{1 - \frac{\lambda_2}{d}x^2\right\} W_{st}(x)$$

$$X_4(x) = A\left\{1 - \frac{\lambda_4}{d}x^2 + \frac{\lambda_4(\lambda_4 - \lambda_2)}{6d^2}x^4\right\} W_{st}(x)$$

$$X_6(x) = A\left\{1 - \frac{\lambda_6}{d}x^2 + \frac{\lambda_6(\lambda_6 - \lambda_2)}{6d^2}x^4 - \lambda_6\frac{(\lambda_6 - \lambda_2)(\lambda_6 - \lambda_4)}{6.15\,d^3}x^6\right\} W_{st}(x)$$

$$\vdots \qquad (23)$$

$$X_1(x) = B\{x\}\, W_{st}(x)$$

$$X_3(x) = B\left\{x - \frac{(\lambda_3 - \lambda_1)}{3d}x^3\right\} W_{st}(x)$$

$$X_5(x) = B\left\{x - \frac{(\lambda_5 - \lambda_1)}{3d}x^3 + \frac{(\lambda_5 - \lambda_1)(\lambda_5 - \lambda_3)}{3.10\,d^2}x^5\right\} W_{st}(x)$$

$$\vdots$$

with corresponding eigenvalues,

$$\lambda_0 = 0 \qquad \lambda_1 = -a$$

$$\lambda_2 = -(c + 2a) \qquad \lambda_3 = -(3c + 3a)$$

$$\lambda_4 = -(6c + 4a) \qquad \lambda_5 = -(10c + 5a) \qquad (24)$$

$$\lambda_6 = -(15c + 6a) \qquad \lambda_7 = -(21c + 7a)$$

$$\vdots \qquad \vdots$$

Since the differential equation for the Cauchy case has only complex singularities in the finite plane, its solution is valid for all finite values of real x, or

$$X_n = \sum_n \phi_n(x)\, W_{st}(x) \qquad (25)$$

where $\phi_n(x)$ are the orthonormal polynomials of the generalized Cauchy p.d.f.

The boundary conditions, Eq. (10), imposed on each X_n will be satisfied depending on the value of m. Specifically, $G[X_n] = 0$ for $x = \pm\infty$ with n such that the moments of the generalized Cauchy p.d.f. are finite. Note that as $c \to 0$ ($h \to \infty$), the set of eigenvalues, Eq. (24), reduces to the eigenvalues of the Gaussian process, and the eigenfunctions are expressed in terms of the Hermite polynomials. Such results should be expected since $K_1(x)$ and $K_2(x)$ would then become the intensity coefficients of the Gaussian process.

D. Evaluating the Autocorrelation Function Using the Differential Equation Technique

The development of the differential equation technique applied to a Markov process may be used to find such important results in the analysis of communication systems as the autocorrelation function at the output of a nonlinearity, the multidimensional p.d.f., the spectral density. The two-dimensional p.d.f. can be expressed in terms of the eigenfunctions found before,

$$W_\tau(x, x_0) = \sum_{j=o}^{\infty} X_j(x)\, X_j(x_0)\, e^{-\lambda_j |\tau|} \tag{26}$$

from which the autocorrelation function can be obtained,

$$R_y(\tau) = \sum_{j=1}^{\infty} h_j^2\, e^{-\lambda_j |\tau|} \tag{27}$$

where the coefficients h_j depend on the nonlinearity used and the eigenfunctions associated with the noise affecting the system,

$$h_j = \int_{-\infty}^{\infty} F(x)\, X_j(x)\, dx \tag{28}$$

For an odd nonlinearity, it is worthwhile to note that the coefficients h_j vanish for even polynomials.

The nonlinearity of Fig. 1 is of a general form. Depending on the values of a_1, a_2 and a_3, such nonlinearity may yield several different limiters frequently used in the analysis of communications systems, such as, the "hard" limiter ($a_2, a_3 \to \infty$), the window limiter ($a_1 = 0$, $a_3 = a_2$), the triangular limiter ($a_2 = a_1$). Furthermore, such nonlinearities, while providing a good approximation to the optimum nonlinearity discussed by Kurz,[2] may be more easily implemented.

Applying the differential equation techniques, a computer program was developed to find a nonlinearity of the form of Fig. 1 which would minimize the power of the noise

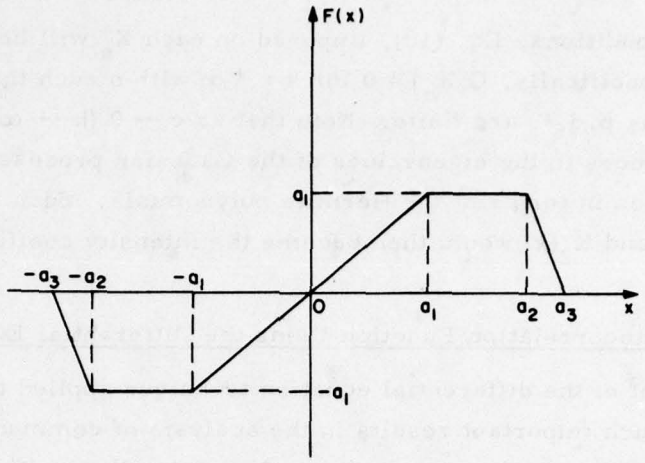


Fig. 1. Nonlinearity used in suboptimum detectors.

$R_y(0)$, under both noise conditions, the Gaussian and the generalized Cauchy noise. Simulations conducted under different noise conditions showed that the output power of the noise increases as a_2 and a_3 increase. It is worth noting here again that the generalized Cauchy p.d.f. used, with m as a parameter, yields the Gaussian p.d.f. ($m \to \infty$), the Cauchy p.d.f. ($m = 1$), with different degrees of noise impulsiveness characterized by different values of m. The triangular nonlinearity yields the smallest $R_y(0)$.

For the linear case, the analysis becomes simple since the coefficients h_j, $j > 1$ in Eq. (27) vanish, yielding for the Gaussian case

$$R_y(\tau) = \sigma^2 e^{\frac{-d}{2\sigma^2}|\tau|} \qquad d > 0 \tag{29}$$

and for the generalized Cauchy distribution

$$R_y(\tau) = \sigma^2 \frac{(m - 1/2)}{(m - 3/2)} e^{-c(m-1)|\tau|} \tag{30}$$

where the parameters m, c, d, σ^2, were described previously. Note that $R_y(0)$ in Eq. (30) decreases as the noise becomes less impulsive (larger m), and reduces to Eq. (29) as $m \to \infty$.

E. Optimizing the Nonlinearity of Fig. 1 Using the Fletcher-Powell Search Procedure

Determining the most efficient nonlinearity of the form of Fig. 1, by varying the parameters a_1, a_2 and a_3 is a brute force method. Instead, a formal search procedure, suitable for the optimization problem, is developed here. The proposed procedure is

a special application of the Fletcher and Powell method of References 3 and 4. It performs the calculation of an unconstrained minimum of a function of several variables. (The procedure of Ref. 3 can also be used to handle constraints on the system.)

The criterion of performance is the minimization of the noise-to-signal ratio, N/S, at the output of the nonlinearity. The case of a constant signal A is considered, with the noise present characterized by the generalized Cauchy p.d.f. The generalized Cauchy p.d.f., Eq. (4), can be written in the case of a shift in the mean as

$$p(x - A) = \frac{K}{((x - A)^2 + h)^m} \tag{31}$$

A program was used to carry out the optimization problem. Using Ref. 3, the gradient functions (partial derivatives of N/S) with respect to a_1, a_2 and a_3 were evaluated.

Basically, the Fletcher-Powell procedure is an iterative process. The direction of the optimal solution is provided depending on the algebraic sign of the gradient functions. The optimum solution is achieved (minimization of N/S) when the gradient functions with respect to a_1, a_2 and a_3 reduce to zero. Under such conditions, the optimum nonlinearity is obtained with the optimal values a_i, $i = 1, 2, 3$ found which minimize N/S.

The results are plotted in Figures 2 through 7. Note again that the signal-to-noise ratio, S/N, was jointly optimized with the nonlinearity, under each noise condition. The simulation results show that a_1, a_2 and a_3 increase as the noise becomes less impulsive, showing an increase in the linear region of the limiter, and providing at the same time less clipping. As the noise becomes less impulsive (larger m or smaller σ^2), the output S/N increases as the signal strength A increases. It leads one to conclude that the nonlinearity performs efficiently under various input S/N ratios.

Joint Services Technical Advisory Committee
F44620-74-C-0056

R. Chassaing and L. Kurz

REFERENCES

1. R.L. Stratonovich, "Topics in the Theory of Random Noise," Vol. I, translated by R.L. Silverman (New York: Gordon and Breach, 1963).
2. L. Kurz, "A Method of Digital Signalling in the Presence of Additive Gaussian and Impulsive Noise," 1962 IRE International Convention Record, pt. 4, Vol. 10, pp. 161-173 (1962).
3. Fletcher and Powell, "Rapidly Convergent Descent Method for Minimization," Computer Journal, Vol. 6, pp. 163-168 (1963).
4. "Scientific Subroutine Package," S.S.P., IBM, Subroutine FMFP.

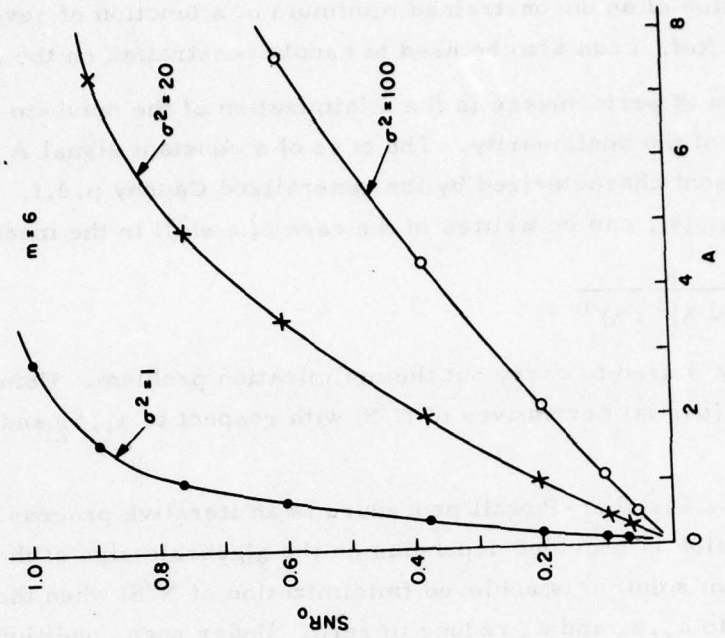
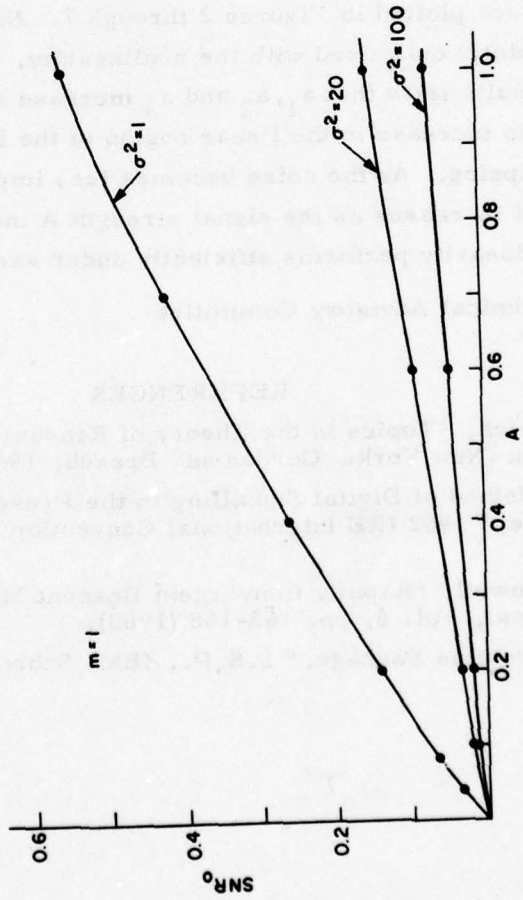


Fig. 3. Output signal-to-noise ratio as a function of signal strength A.



Fif. 2. Output signal-to-noise ratio as a function of signal strength A.

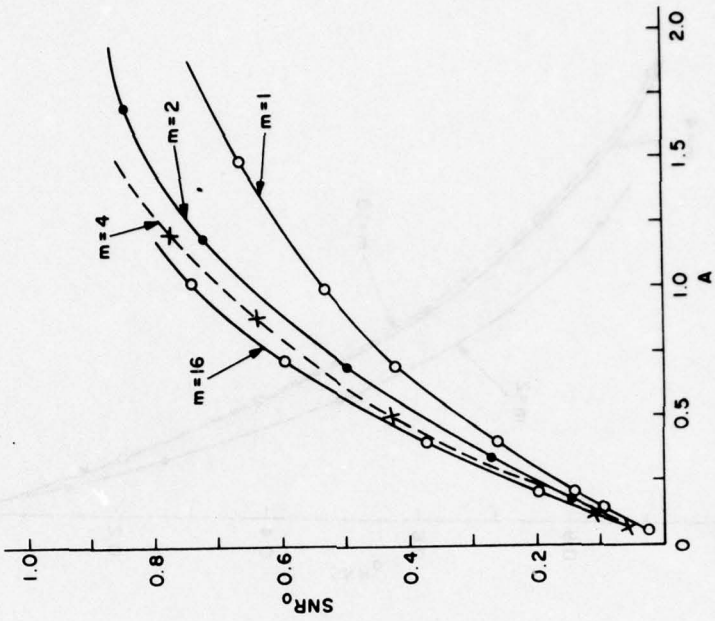


Fig. 5. Output signal-to-noise ratio as a function of signal strength A.

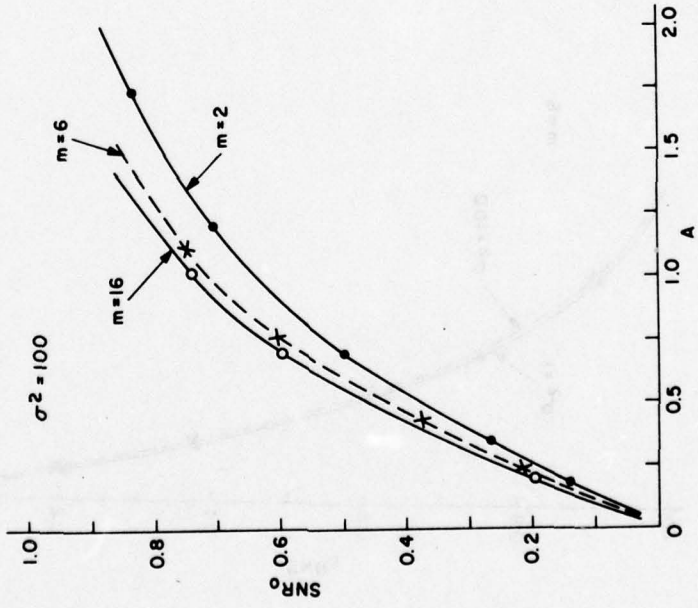


Fig. 4. Output signal-to-noise ratio as a function of signal strength A.

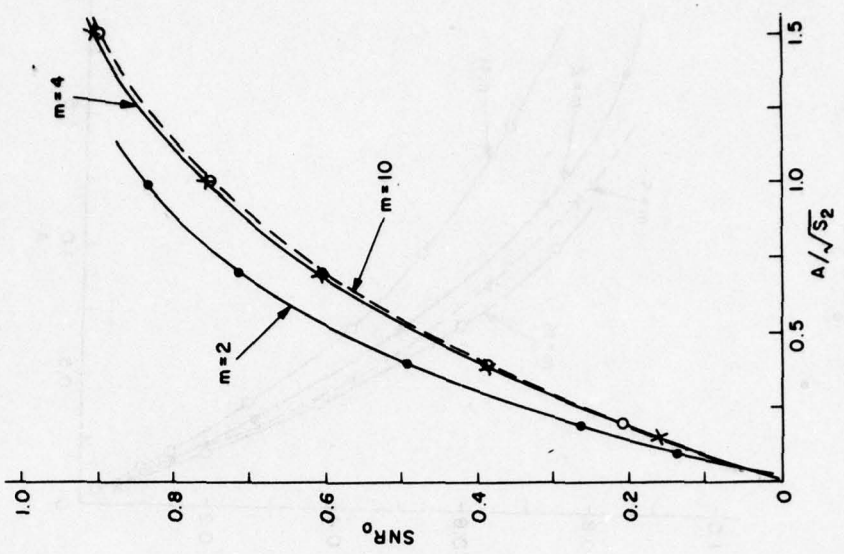


Fig. 7. Output signal-to-noise ratio as a function of input signal-to-noise ratio.

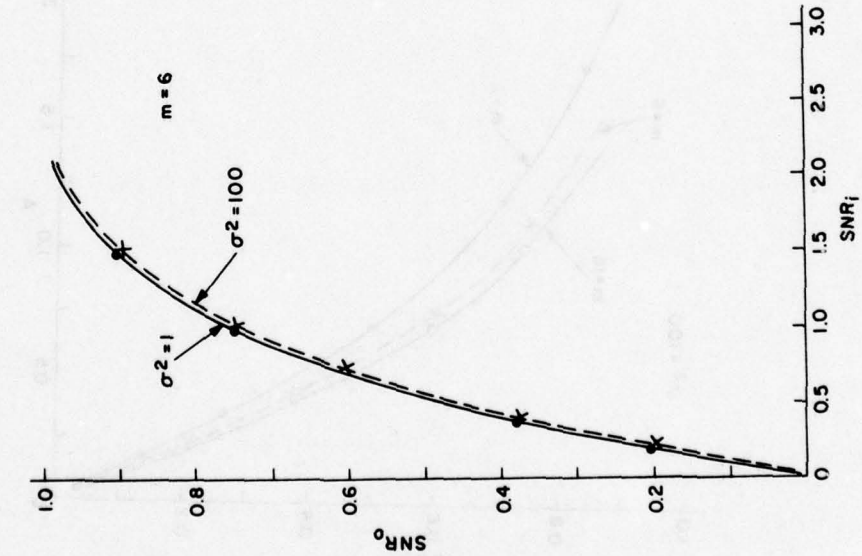


Fig. 6. Output signal-to-noise ratio as a function of input signal-to-noise ratio.

A UNIQUE FREQUENCY DIVISION MULTIPLEX COMMUNICATION SYSTEM*

L. Kurz and J. Reed

In this report, a brief description of a unique FDM system is given. Details of its performance and a comparison study with an existing system are given in Reference 1.

For the actual FDM system, basic parameters and architecture for a telecommunications switching system using radio frequency carriers on a coaxial cable in a ring main form are developed. The resulting distributed system achieves transmission by means of single sideband modulation. It also achieves four wire connectibility by transmitting either the main or image signal generated in mixers, using the fact that the information passband is limited and filtering may be applied.

Similar systems have been proposed and built previously. In both cases dialing is accomplished by frequency setting using digitally controlled synthesizers. Such a system had a receiving system using direct (synchronous) conversion directly to audio. This report proves that such a restriction is not necessary and that the same or better results can be achieved by using a more general formulation involving both transmitting and receiving intermediate frequencies different from zero.

A further result of removing the restriction and receiving IF is that it becomes readily possible to use independent sideband transmission. This provides signalling channels for improving the system performance, its connectibility and flexibility.

This report compares both of these systems by developing the frequency plan for each and demonstrating that the original system is a special case of the one developed herein. The system architecture is demonstrated in the form of schematics and block diagrams.

An investigation is made of the resulting system to develop further the design rules and to demonstrate the practicality of the proposed scheme. Sources of noise are enumerated and examined, in some detail, in order to show the limits presently achievable in signal-to-noise performance. Intermodulation products are enumerated and evaluated by the use of a modified form of Bennet's method.

Formulas are derived for choosing frequencies based upon the number of subscribers requiring service and the achievable transmission bandwidth. Examples of the use of these formulas are given showing the limitations on intermodulation terms generated in terms of cross-over products.

Sensitivity calculations are made to demonstrate the absence of need for broadband amplifiers with their unavoidable nonlinearities, leading to poorer IM performance.

---

*Patents involving the system summarized here are pending.

Power outputs from transmitters are shown to be those normally used in FDM systems. Sensitivity calculations account for cable losses, coupler losses, power loss due to multitude subscribers and to cable terminations.

The special advantages of the more general system described is elucidated by an examination of the signalling and supervisory problem which was not satisfactorily solved in the previous system. The conferencing, override and pre-empt problems posed by this type of switch are each examined and several alternate solutions to these problems are proposed.

In each case the solution to the problems of signalling are clarified by proposing algorithms for designing the logic system to be used in making the various forms of connections. These algorithms are given in terms of single channel dialing sequences. Discussions of problems which arise in these sequences are provided together with proposed methods of overcoming each problem. These solutions are further clarified by means of block diagrams demonstrating proposed subscriber sets.

Indications are given where further profitable investigations both theoretic and experimental are needed.

The conclusion drawn is that the general system formulas lead to a superior design solving several previously unsolved problems.

ITT - Military Electronic System L. Kurz and J. Reed

REFERENCE

1. J. Reed, "A Unique Frequency Division Multiplex Communication System," Ph.D. Dissertation, Polytechn. Inst. of New York (June 1977).

ON CHANNEL MONITORING TECHNIQUES IN THE PRESENCE OF IMPULSIVE NOISE OR FADING EFFECTS

I. M. Habib and L. Kurz

The robust p. d. f. estimation procedures presented by the authors in the previous report[1] are used to obtain channel monitoring estimators which are used to identify whether the noise in the communication channel is Gaussian, a mixture of Gaussian and impulsive noise, or the channel operates in a fading mode. The estimators presented here are helpful in developing adaptive detectors in a non-Gaussian noise environment.

A. Some Representations of Impulsive Noise

Impulsive noise is generally characterized by pulses of short duration and large amplitude which corrupts, to a certain extent, almost every transmission path. Therefore, it is important that detection procedures be optimized to detect a signal corrupted by a mixture distribution. Such procedures were developed in the literature based on various mathematical representations of the impulsive noise.

This section is aimed to serve as a review of the various impulsive noise representations used in the literature.

An impulsive noise model that has been used successfully, assumes that the impulsive noise consists of very narrow pulses (approximating impulses) occurring randomly, where the frequency of pulse occurrence is governed by certain distribution. This model was used in the early work of Kurz[2] for detecting a signal in additive Gaussian and impulsive noise, with the frequency of pulse occurrence assumed to follow a Poisson distribution

$$f(n, t) = \frac{(aT)^n}{n!} e^{-aT} \tag{1}$$

where n is the number of pulses in an interval T and a is the long term average of the number of pulses occurring per unit time. In this model, the distribution of amplitudes was assumed to be of large variance but otherwise unspecified. The Poisson distribution of occurrence of noise impulses can be easily generalized by replacing Eq. (1) with a type B Gram-Charlier expansion.

A different representation of the impulsive noise assumes that the noise process is represented by an independent increments process, where the p. d. f. associated with the sample amplitude of the process may be given by a Cauchy amplitude distribution, whose p. d. f. is given by

$$f(x) = \frac{\lambda}{\pi} \frac{1}{x^2 + \lambda^2} \tag{2}$$

where λ is a specifying parameter[3]. This model was also used by Kurz and Kapp[4] for evaluating the performance of suboptimum detectors in Gaussian and impulsive noise. In general, this model is used when the impulsive noise is severe.

Alternately, an exponential amplitude distribution, with p. d. f. given by

$$f(x) = ae^{-b|x|^n} \qquad -\infty < x < \infty \tag{3}$$

where a, b and n, specify the impulsiveness of the noise, is a useful representation for a large class representing a weak impulsive noise.

B. Gaussian and Weak Impulsive Noise Channel Estimator

The communication channel in this section is assumed to be corrupted by Gaussian and weak impulsive noise. The Gaussian noise is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{4}$$

where μ and σ are the average and standard deviation of the distribution respectively.

The weak impulsive noise belongs to the exponential class and is given by the double-exponential distribution*

$$f(x) = \frac{\alpha}{2} e^{-\alpha|x|} \qquad -\infty < x < \infty \tag{5}$$

where α is a specifying parameter of the distribution. The distribution of the noise in the channel is a mixture of the Gaussian and impulsive noise with the c. d. f. given by

$$F(x) = (1 - \delta)\, F_1(x) + \delta\, F_2(x) \qquad \infty - < x < \infty \tag{6}$$

where $F(x)$ is the c. d. f. of the mixture distribution
$F_1(x)$ is the c. d. f. of the Gaussian noise
$F_2(x)$ is the c. d. f. of the impulsive noise
δ is the mixing parameter.

For this class of communication channels, the noise p. d. f. will be represented by a series expansion of the orthogonal Hermite functions, i. e.,

$$\hat{f}(n(t_j)) = \sum_{i=1}^{r} a_i \phi_i(n(t_j)) \tag{7}$$

*Similarly, one may proceed with the more general distribution with the amplitude p. d. f. given by Equation (3).

where a_i $(i=1,\ldots\ldots\ldots,r)$ are the expansion coefficients

$\phi_i(u)$, $(i=1,\ldots\ldots\ldots,r)$ are the Hermite functions given in Reference 1

$n(t_j)$ are the noise samples (the output of a sampler)

The expansion coefficients are given by

$$a_i = \int_u \phi_i(u) f(u) du \tag{8}$$

A recursive estimation procedure for estimating the a_i's using the Robbins-Monro procedure has been developed in Reference 1.

The resulting recursive equation for estimating a_i is given by

$$a_i^{(k+1)} = a_i^{(k)} + \gamma_k \left[\phi_i(n(t_k)) - a_i^{(k)} \right] \tag{9}$$

where γ_k is a gain coefficient of the form $1/k$

As the number of the processed noise samples, $n(t_k)$, increases, $a_i^{(k)}$ approaches a_i leading to a better estimator. This noise estimator is actually the same estimator presented in Ref. 1 for the Gaussian class where the random variable X_n is replaced here by the random noise sample $n(t_k)$.

A block diagram that is used to implement this estimator is shown in Figure 1. The noise process in the channel is sampled every Δt seconds and the samples are processed by the p. d. f. estimator.

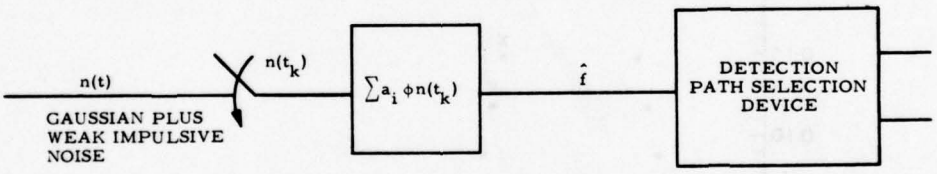


Fig. 1. Channel noise estimator for channels with Gaussian and/or weak impulsive noise.

The sampling time Δt is chosen based on the type of transmitted signals. For example, if the signals transmitted consist of a sequence of pulses, each of duration T, then the sampling may be made over a time period (10T) at a rate (T/100) seconds. To ensure that the samples used in the estimation procedure are independent, the samples generated from the first time period are interleaved with those generated in the second, third,... and tenth time periods. This interleaving results in samples that are i. i. d. and suitable for processing by the p. d. f. estimator.

Upon identifying the channel noise, the detection path decision device selects the appropriate detection procedure. It is to be noted that this estimator is adaptive

to the type of noise in the channel without changing any parameters of the estimator, a particularly useful characteristic in practical systems. The results of computer simulation of this estimator for a mixture of (0, 1) Gaussian noise and a double-exponential distribution ($\alpha = .5$) is shown in Figs. 2, 3 and 4 for mixing parameter values $\delta = 0$, .1, and 1 respectively. The range of the mixing parameter values used in the simulation allow the examination of the performance of the estimator under all noise conditions assumed in the channel. The graphs in Figs. 2, 3 and 4 show the closeness of the noise estimate to the exact distribution and therefore its practical usefulness. However, for a purely impulsive noise ($\delta = 1$), more samples are required to estimate the cusp of the p. d. f.

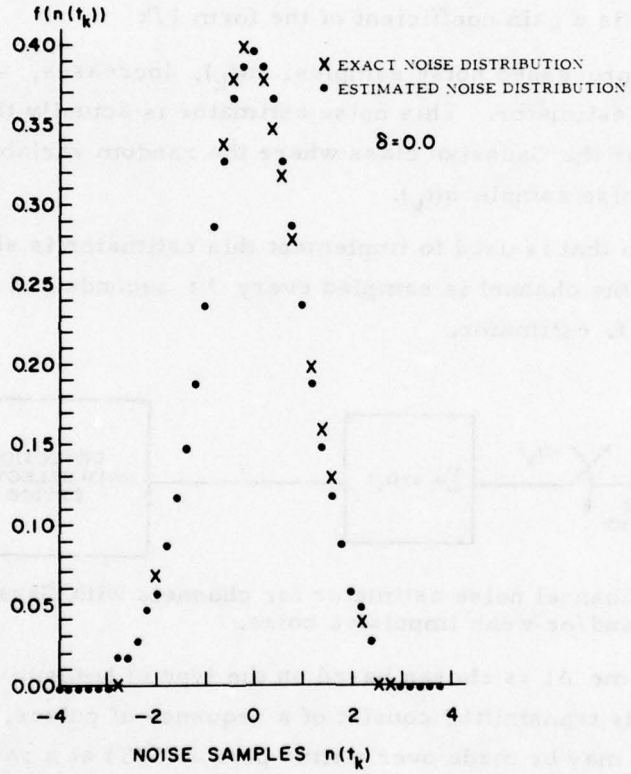


Fig. 2. Gaussian and weak impulsive noise channel estimator.

This estimator, in other words, is a robust estimator for the class of communication channels corrupted by a mixture of Gaussian and weak impulsive noise, where robustness is defined in the sense of Reference 1.

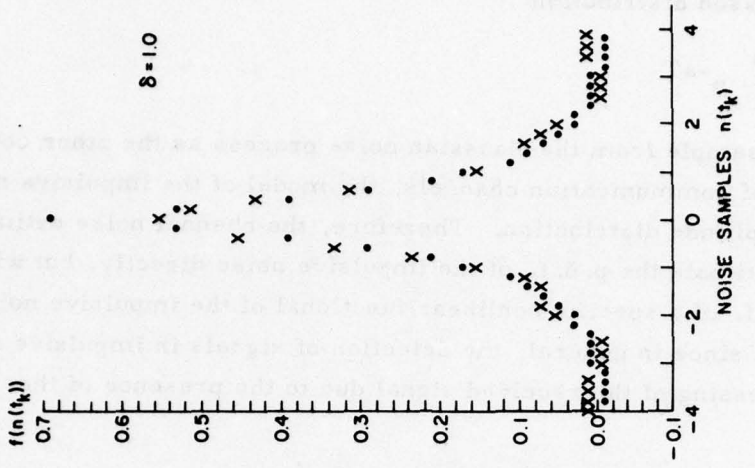


Fig. 4. Gaussian and weak impulsive noise channel estimator.

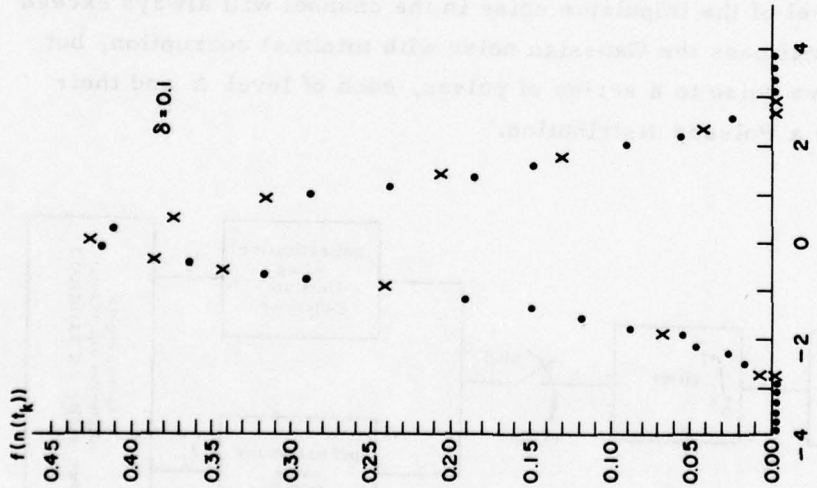


Fig. 3. Gaussian and weak impulsive noise channel estimator.

C. Gaussian and Severe Impulsive Noise Channel Estimator

The communication channel considered in this section is assumed to be corrupted by a mixture distribution with severe impulsive noise whose frequency of occurrence in time is given by a Poisson distribution

$$f(n, t) = \frac{(aT)^n}{n!} e^{-aT} \tag{10}$$

as one component and a sample from the Gaussian noise process as the other component. For this class of communication channels, the model of the impulsive noise does not assume any amplitude distribution. Therefore, the channel noise estimator for this class will not estimate the p. d. f. of the impulsive noise directly, but will rather estimate the p. d. f. of a specific nonlinear functional of the impulsive noise. This is especially useful since in general, the detection of signals in impulsive noise requires nonlinear processing of the received signal due to the presence of the impulsive noise.

A nonlinear device that will be used for processing the received noise is shown in Figure 5. This device consists of a soft limiter followed by an integrator, and a sampler. The sampled noise is processed simultaneously by two p. d. f. estimators followed by a threshold decision device. It is assumed that the received Gaussian noise level will not exceed the level A. This level could be adjusted periodically based on the estimation of the Gaussian noise variance, σ^2, in the channel and setting $A = 4\sigma$. In addition, the level of the impulsive noise in the channel will always exceed A. The nonlinear device will pass the Gaussian noise with minimal corruption, but will transform the impulsive noise to a series of pulses, each of level A and their occurrence in time follows a Poisson distribution.

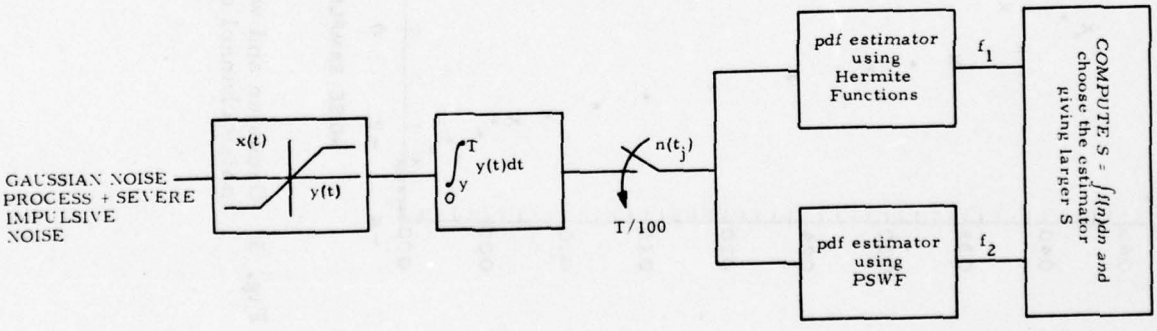


Fig. 5. Channel noise estimator for channels with Gaussian and/or severe impulsive noise.

The output of the nonlinear device is integrated over a time period T which could be selected according to the type of signals transmitted. For example, in digital transmission, T could be chosen as the inverse of the pulse rate.

In summary, the operation of the soft-limiter-integrator is as follows:

An integrator for the Gaussian noise.

A hard-limiter followed by an integrator for the impulsive noise.

In impulsive noise, the p. d. f. of the detector integrator is the same class as in Reference 1. The p. d. f. of the amplitude of the clipped and integrated impulsive noise is given by

$$f(Y_T) = \frac{ae^{-kT}}{2}\left[I_0\left(k\sqrt{1 - Y_T/T)^2}\right) + \frac{I_1\left(k\sqrt{1 - (Y_T/T)^2}\right)}{\sqrt{1 - (Y_T/T)^2}}\right] + \frac{e^{-k}}{2}\left[\delta(Y_T - T) + \delta(Y_T + T)\right] \quad |Y_T| < T \qquad (11)$$

$$= 0 \qquad |Y_T| > T$$

where

I_0 and I_1 are the modified Bessel functions of the first kind, order zero and one, respectively,

and

$k = aT$.

In Gaussian noise, the output of the integrator is a random variable whose average and standard deviation are functions of the mean, the autocorrelation function of the input Gaussian process and the integration period T.

On the assumption that the noise is additive, the output of the integrator is given by

$$n(t) = (1 - \delta)\, n_1(t) + \delta\, n_2(t)$$

where

$n_1(t)$ is a (μ, σ^2) Gaussian noise

$n_2(t)$ is impulsive noise whose p. d. f. is given by Eq. 11

δ is the mixing parameter.

The output of the integrator is sampled every Δt seconds where Δt and the sampling

period are chosen as in Section B to ensure independency between samples. The resulting samples are then used to estimate the p. d. f. of the noise in the channel.

The noise samples are processed by two different p. d. f. estimators. The first estimator is of the form

$$\hat{f}(n(t_j)) = \sum_{i=1}^{r} a_i \phi_i(n(t_j)) \tag{12}$$

where

$a_i (i=1, \ldots, r)$ are the expansion coefficients

and

$\phi_i(u)$ $(i=1, \ldots, r)$ are the Hermite functions.

The coefficients a_i are given by

$$a_i = \int_u \phi_i(u) f(u) du \tag{13}$$

and are estimated recursively according to the following equation

$$a_i^{(k+1)} = a_i^{(k)} + \gamma_k \left[\phi_i(n(t_k)) - a_i^{(k)} \right] \tag{14}$$

The second estimator is of the form

$$\hat{f}(n(t_j)) = \sum_{i=1}^{l} b_i \psi_i(n(t_j), c) \tag{15}$$

where

b_i $(i=1, \ldots, l)$ are the expansion coefficients

$\psi_i(u, c)$, $(i=1, \ldots, l)$ are the prolate spheroidal wave functions.

The expansion coefficients b_i's are given by

$$b_i = \int_u f(u) \psi_i(u, c) du \qquad (i=1, \ldots, l) \tag{16}$$

These coefficients are estimated recursively using the Robbins-Monro procedure. The recursive equation is given by

$$b_i^{(k+1)} = b_i^{(k)} + \gamma_k \left[\psi_i(n(t_k)) - b_i^{(k)} \right] \tag{17}$$

As the number of the processed noise samples, $n(t_k)$, increases, the estimate $b_i^{(k)}$

approaches b_i and a better p. d. f. estimation is obtained. The choice of the proper noise p. d. f. is based on calculating the statistic S given by

$$S = \int \hat{f}(u)du \tag{18}$$

for each of the estimators and choosing the one which gives a larger S.

This procedure was simulated for a mixture of a severe impulsive and Gaussian noise where the output of the integrator is given by Equation (11). The results of the computer simulation is shown in Figs. 6, 7 and 8 for various mixing parameters $\delta = 0, .1, 1$. The values of the statistic S for the various δ's are shown in Table I. These results indicate the good performance of the statistic S as a channel condition estimator, especially in the limiting cases and that this estimator performs well under any mixture of Gaussian and severe impulsive noise. For a pure impulsive noise, the estimator of the form Eq. (15) is the proper one. For a pure Gaussian noise, this estimator has very poor performance, however, the proper p. d. f. is given by the estimator of the form Equation (12). For mixing parameter values between 0 and 1, the better estimator depends on the dominant noise. The statistic S will be the basis of the selection of one estimator versus the other.

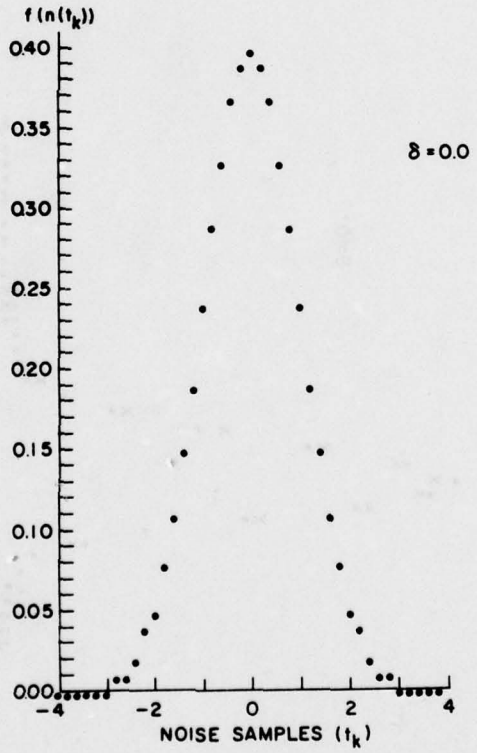


Fig. 6. Gaussian and severe impulsive noise channel estimator.

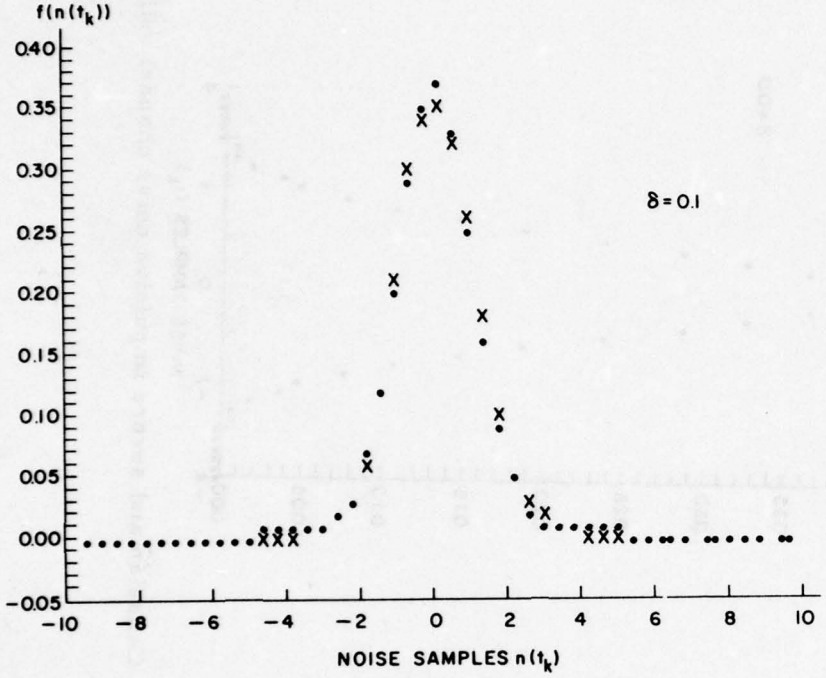


Fig. 7. Gaussian and severe impulsive noise channel estimator.

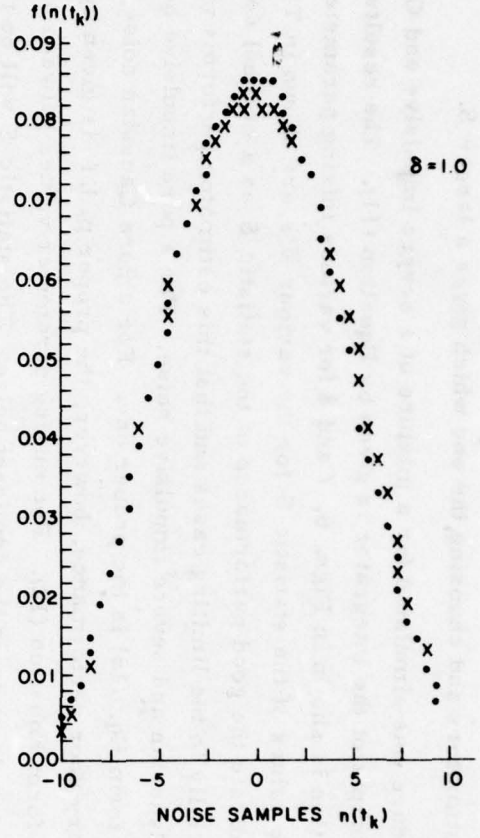


Fig. 8. Gaussian and severe impulsive noise channel estimator.

TABLE I. Values of the statistic S for various mixtures of Gaussian and impulsive noise.

	S	
	Hermite functions	P. S. W. F.
0	0. 999	0. 92
0. 1	0. 968	0. 92
1. 0	0. 655	0. 99

A different approach for selecting the proper orthogonal set is based on estimating the first coefficient in both the series expansion, Equations (12) and (15). The resulting estimates, a_1 and b_1, are compared and the orthogonal set yielding a higher value for the first term is selected for estimating the noise p. d. f. This procedure is justified on the basis that if the proper orthogonal set for the underlying distribution is selected, then the first term in the series expansion contributes most to the estimated p. d. f. Higher order terms have less contribution and may be considered to have only a smoothing effect. This procedure is attractive because it is less expensive than the previous one and may be easily implemented in a practical communication system.

Figure 9 shows a block diagram that implements this procedure for estimating the channel noise condition. The results of estimating the first coefficient for a Gaussian and impulsive noise using both the P. S. W. F. and the Hermite sets is shown in Table II.

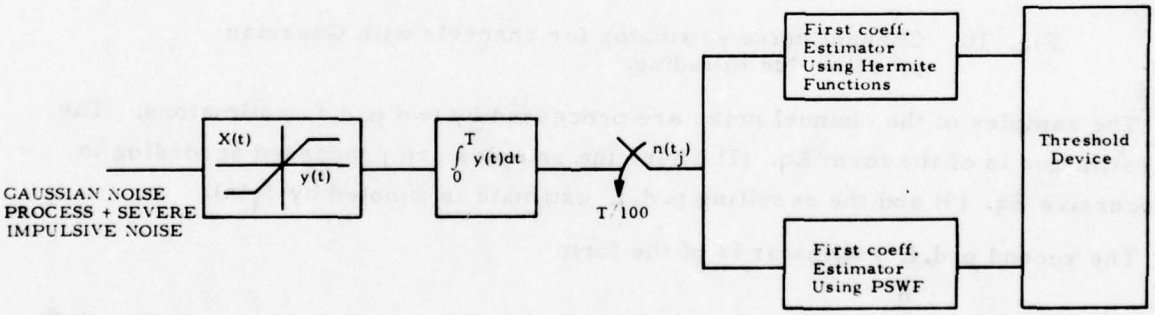


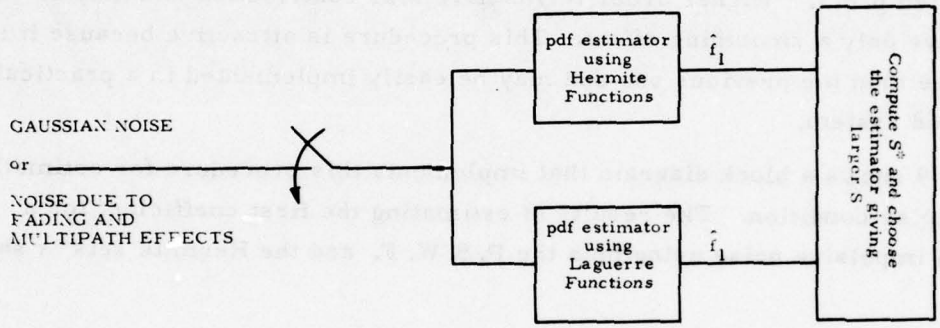
Fig. 9. Channel noise estimator for channels with Gaussian and/or severe impulsive noise

TABLE II.

Distribution	using Hermite functions	using P. S. W. F.
Gaussian noise	0.531	0.49
Impulsive noise	0.14	0.2447

D. Gaussian and Fading Channel Noise Estimator

The communication channel in this section is assumed to be in one of two states, G or B. In the G state, it is assumed that the channel is corrupted by Gaussian noise of the form Equation (4). In the B state, the channel is assumed to be in a fading state with amplitude distribution of the one-sided class (e. g. lognormal, Rayleigh or Beta). A block diagram of the estimator that will be used in identifying the channel noise condition is shown in Figure 10.



*$S = \int f(n)\,dn$

Fig. 10. Channel noise estimator for channels with Gaussian or noise due to fading.

The samples of the channel noise are processed by two p. d. f. estimators. The first estimator is of the form Eq. (7) where the samples are processed according to the recursive Eq. (9) and the resulting p. d. f. estimate is denoted by $f_1(n)$.

The second p. d. f. estimator is of the form

$$\hat{f}(n(t_j)) = \sum_{i=1}^{q} c_i L_i(n(t_j)) \tag{19}$$

where

$c_i (i=1,\ldots,r)$ are the expansion coefficients

$L_i(u)$ $(i=1,\ldots.r)$ are the Laguerre polynomials defined in Reference 1

$n(t_j)$ are the noise samples (obtained at the output of the sampler).

The coefficients c_i are given by

$$c_i = \int_u f(u)L_i(u)du \tag{20}$$

A recursive procedure has been developed for this class of estimators in Ref. 1 to estimate the coefficients c_i. The resulting recursive equation is given by

$$C_i^{(k+1)} = c_i^{(k)} + \gamma_k\left[L_i(n(t_k)) - c_i^{(k)}\right] \tag{21}$$

where

γ_k is a gain coefficient of the form $1/k$

The second p. d. f. estimator processes the samples according to Eq. (21) and the resulting p. d. f. estimate is denoted by $\hat{f}_2(n)$.

The statistic S defined by

$$S_j = \int_u \hat{f}_j(u)du \tag{22}$$

where j=1 or 2, is calculated for the first (j=1) estimator and the second (j=2) estimator. The p. d. f. estimate leading to the larger statistic is chosen to be the proper p. d. f. The results of simulating the above procedure on the computer for a channel whose G state is represented by a (0, 1) Gaussian noise and the B state is represented by a beta distribution given by

$$f(x) = \begin{cases} \frac{(\alpha+\beta+1)!}{\alpha!\ \beta!}\, x^{\alpha}(1-x)^{\beta} & 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases} \tag{23}$$

where α and β are the specifying parameters of the distribution is shown in Figures 11 and 12.

The values of the statistic S for both states, G and B, are shown in Table III.

The results of this simulation show the excellent performance of the channel noise estimator and the sensitivity of the statistic S for the proper channel noise estimate.

A different approach to estimation of the channel noise conditions is based on estimating the first term in the p. d. f. series expansion using the Hermite and the

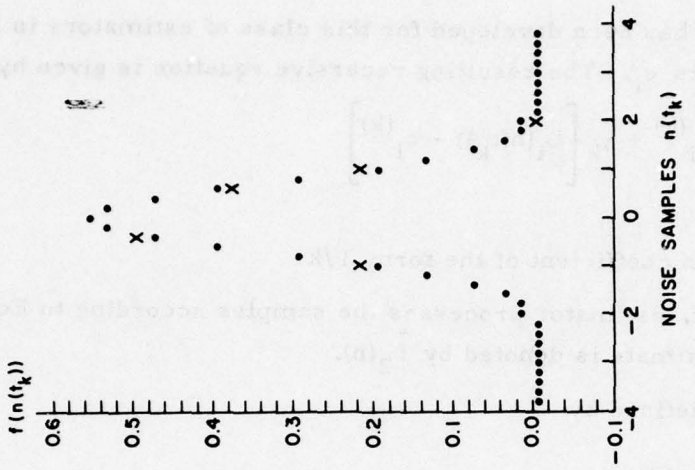

f(n(t_k))
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-4
-2
0
2
4
NOISE SAMPLES n(t_k)

Fig. 12.

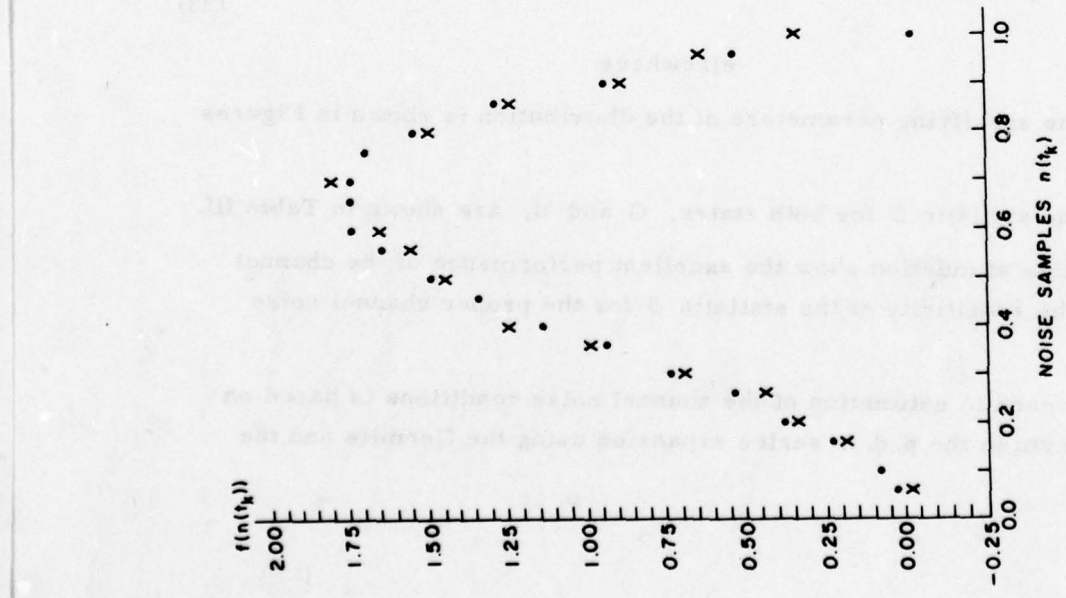

f(n(t_k))
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
-0.25
0.0
0.2
0.4
0.6
0.8
1.0
NOISE SAMPLES n(t_k)

Fig. 11.

TABLE III. Values of the statistic S for Gaussian noise or noise due to fading.

Distribution	S: Hermite functions	S: Laguerre functions
Gaussian noise	0.999	0.621
Noise due to fading (from the class of one-sided dist.)	0.533	0.932

Laguerre polynomials. The estimated terms, a_1 and c_1, are compared and the set yielding the higher value for the first term is used to estimate the noise p.d.f. Figure 13 shows a block diagram that implements this procedure. The results of estimating a_1 and c_1 for a Gaussian noise and noise in a fading channel is shown in Table IV.

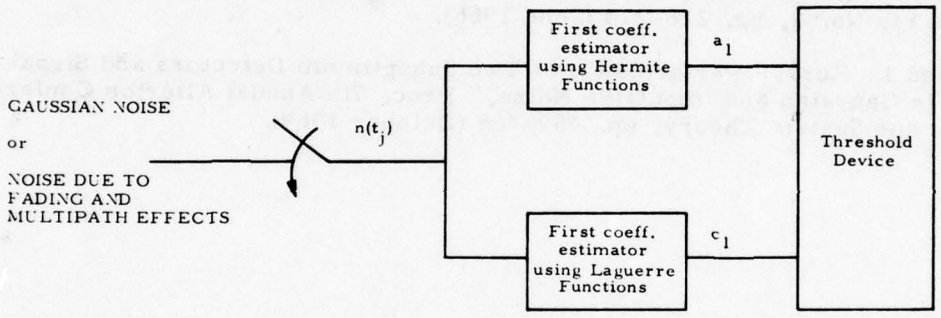


Fig. 13. Channel noise estimator for channels with Gaussian or noise due to fading.

TABLE IV.

Distribution	The value of the first term in the expansion using Hermite functions	using Laguerre polynomials
Gaussian noise	0.531	0.314
Noise due to fading	0.23	0.697

The introduction of the previous three channel noise condition estimators is intended to show the application of the p.d.f. estimator developed in Chapter 3 to the

field of communication engineering. The specific problems considered are intended to be used in the detection of signals corrupted by Gaussian and/or non-Gaussian noise (e. g. impulsive or amplitude fading). The results of these applications show the usefulness of these procedures for channel noise identification and the sensitivity of the statistic S to the proper noise model.

Joint Services Technical Advisory Committee
F44620-74-C-0056

I. M. Habib and L. Kurz

RCA Globcom

REFERENCES

1. I. M. Habib and L. Kurz, "Nonparametric Probability Density Estimation," Progress Report No. 41 to JSTAC, Polytechn. Inst. of New York, Report No. R452. 41-76, pp. 586-591 (1976).

2. L. Kurz, "A Method of Digital Signalling in the Presence of Additive Gaussian and Impulsive Noise," IRE International Convention Record, Vol. 10, p. t. 4 (1962).

3. S. S. Pappaport and L. Kurz, "An Optimal Nonlinear Detector for Digital Data Transmission through Non-Gaussian Channels," IEEE Trans. on Com. Tech., Vol. Com-14, No. 3, pp. 226-274 (June 1966).

4. J. Kapp and L. Kurz, "Performance of Two Suboptimum Detectors and Signal Selection in Gaussian and Impulsive Noise," Proc. 7th Annual Allerton Conference on Circuit and System Theory, pp. 459-466 (October 1969).

SUPPRESSION OF TIMING JITTER IN PARTIAL RESPONSE SYSTEMS

M. Kavehrad and L. Kurz

The rise in demand for rapid data transmission through telephone lines led to major research activity in improving equalization and general performance of PAM systems. In particular, numerous papers have been written about the partial response systems. The effort here is concentrated on the study of partial response systems suggested by Kretzmer[1] if the timing jitter immunity constraints are included. In particular, the timing jitter suppression is achieved by imposing the zero-derivative constraint at all sampling points. It is shown that a signal of minimum bandwidth, satisfying zero-intersymbol interference and zero-derivative constraints, is twice the Nyquist bandwidth. It is proven that the overall system response in the frequency domain is triangular followed by a tapped delay line. In addition, properties of signals which exceed twice the Nyquist rate are explored. Modifications of the Kretzmer's signals including the timing jitter suppression are given.

A. Mathematical Formulation of the Problem

Following Gibby and Smith,[2] the system response in time r(t) and in frequency $R(\omega) = A(\omega)\, e^{j\alpha(\omega)}$ are related, if Nyquist conditions are satisfied, by

$$r_k = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \int_{-\pi/T}^{\pi/T} R(u + \frac{2n\pi}{T})\, e^{jukT}\, du \tag{1}$$

where

$$\omega = u + \frac{2n\pi}{T} \quad \text{and} \quad r_k = r(kT),\ k = 0, 1, 2, \cdots$$

Introducing the notation

$$R_n = R(u + \frac{2n\pi}{T}),\ A_n = A(u + \frac{2n\pi}{T}) \quad \text{and} \quad \alpha_n = \alpha(u + \frac{2n\pi}{T}),$$

the Nyquist problem[2] reduces to

$$\sum_{n=-\infty}^{\infty} A_n \cos \alpha_n = T \sum_k r_k \cos ukT \tag{2}$$

$$\sum_{n=-\infty}^{\infty} A_n \sin \alpha_n = -T \sum_k r_k \sin ukT \tag{3}$$

Thus, signals satisfying Eqs. (2) and (3) eliminate intersymbol interference at sampling points $k = 1, 2, 3, \cdots$.

To suppress timing jitter, it is suggested that

$$\frac{d\,r(t)}{dt}\Big|_{t=kT} = 0,\ k = 0, 1, 2, \cdots$$

This condition assures flatness of signals in the neighborhood of the sampling points, yielding immunity to timing jitter. Using similar analysis to the one suggested by Gibby and Smith,[2] suppression of timing jitter add two additional design equations to Eqs. (2) and (3), namely,

$$\sum_{n=-\infty}^{\infty} n A_n \cos \alpha_n = \left(\frac{T^2}{2\pi}\right) u \sum_k r_k \cos(kTu) \tag{4}$$

$$\sum_{n=-\infty}^{\infty} n A_n \sin \alpha_n = \left(\frac{T^2}{2\pi}\right) u \sum_k r_k \sin(kTu) \tag{5}$$

We seek a signal of minimum bandwidth which satisfies Equations (2) to (5). If one claims that there exists a signal of bandwidth $n\pi/T$ with $n < 2$ satisfying the four equations, one arrives at a contradiction. Referring to Eqs. (2) through (5), for $n < 2$ only the term corresponding to $n = 0$ contributes to the summations and Eqs. (4) and (5) cannot be satisfied for $0 \leq u \leq (2 - n)\pi/T$. Consequently, the bandwidth corresponding to at least two Nyquist bandwidths is required ($n = 2$). For the two terms ($n = -1$ and 0) contributing to the sums in the interval $0 \leq u \leq \pi/T$, Eqs. (2) to (5) reduce to

$$A_0 \cos \alpha_0 + A_{-1} \cos \alpha_{-1} = T \sum_k r_k \cos kuT \tag{6}$$

$$A_0 \sin \alpha_0 + A_{-1} \sin \alpha_{-1} = -T \sum_k r_k \sin kuT \tag{7}$$

$$-A_{-1} \cos \alpha_{-1} = -\frac{T^2}{2\pi} u \sum_k r_k \cos kuT \tag{8}$$

$$-A_{-1} \sin \alpha_{-1} = \frac{T^2}{2\pi} u \sum_k r_k \sin kuT \tag{9}$$

Solving the above system of equations, one obtains

$$R(\omega) = T\left(1 - \frac{T}{2\pi}|\omega|\right) \sum_k r_k e^{-j\omega kT} \qquad 0 \leq |\omega| \leq \frac{2\pi}{T} \tag{10}$$

which represents a triangular spectrum followed by a tapped delay line. For the standard PAM system, Eq. (10) reduces to the triangular spectrum. For the duobinary system suggested by Lender,[3] Eq. (10) reduces to

$$R(\omega) = \frac{T}{2}\left(1 - \frac{T}{2\pi}|\omega|\right)\cos\frac{\omega T}{2} \qquad 0 \leq \omega \leq \frac{2\pi}{T} \tag{11}$$

The corresponding signal is then

$$r(t) = \frac{1}{4}\left\{\left[\frac{\sin[\frac{\pi}{T}(t+\frac{T}{2})]}{\frac{\pi}{T}(t+\frac{T}{2})}\right]^2 + \left[\frac{\sin[\frac{\pi}{2}(t-\frac{T}{2})]}{\frac{\pi}{T}(t-\frac{T}{2})}\right]^2\right\} \tag{12}$$

B. Signals of Bandwidth Greater Than Twice the Nyquist Bandwidth

Consider a system with signalling contained in the interval $(-n\pi/T, n\pi/T)$, where $2 < n < 3$. Following the procedure outlined in Section A, Eq. (10) is replaced by

$$R(\omega) = \begin{cases} \frac{T}{2}(1 - \frac{T}{2\pi}\omega)\cos(\frac{\omega T}{2}) & \frac{(n-2)\pi}{T} < \omega < (4-n)\frac{\pi}{T} \\ \frac{T}{2}(1 + \frac{T}{2\pi}\omega)\cos(\frac{\omega T}{2}) & -\frac{(4-n)\pi}{T} < \omega < -(n-2)\frac{\pi}{T} \\ \text{undetermined} & \text{otherwise} \end{cases}$$

If, in a data transmission system, we transmit simultaneously with high- and low-speed rate, the low-speed part of the data stream would be contained in $|\omega| < (n-2)\pi/T$. The high-speed part of the data stream will utilize only twice the Nyquist bandwidth. For this type of system, Eq. (11) is replaced by

$$R(\omega) = \begin{cases} 0 & 0 \leq \omega \leq \frac{(n-2)}{T}\pi \\ \frac{T}{2}(1 - \frac{T}{2\pi}|\omega|)\cos(\frac{\omega T}{2}) & (n-2)\pi/T < |\omega| < (4-n)\pi/T \\ \frac{T}{4}(2 - \frac{T}{2\pi}|\omega|)\cos(\frac{\omega T}{2}) & (4-n)\pi/T < |\omega| < n\pi/T \end{cases} \tag{13}$$

Consider partial response signals with $2 < n < 3$. We seek the form of the partial response signals which will suppress timing jitter and utilize minimum energy outside the basic range of frequencies, or the solution to Eqs. (6) to (9) subject to energy constraints and $2 < n < 3$ is our goal.

After some simple mathematical manipulations, the signal constraint may be written as

$$E = E_0 + \frac{1}{\pi}\int_0^{(n-2)\pi/T}[A_0^2(u) + A_1^2(u) + A_{-1}^2(u)]\,du \tag{14}$$

where E_0 represents the fixed energy in the interval $(n-2)\pi/T < \omega < (4-n)\pi/T$.

Using the standard methods of the calculus of variations, the energy constrained solution to Eqs. (6) to (9) is

$$R(\omega) = \begin{cases} \frac{T}{6}\cos(\frac{\omega T}{2}) & 0 \leq |\omega| \leq (n-2)\pi/T \\ \frac{T}{2}(1 - \frac{T}{2\pi}|\omega|)\cos(\frac{\omega T}{2}) & (n-2)\pi/T < |\omega| < (4-n)\pi/T \\ \frac{T}{12}(5 - \frac{3T}{2\pi}|\omega|)\cos(\frac{\omega T}{2}) & (4-n)\pi/T < |\omega| \leq n\pi/T \end{cases} \quad (15)$$

C. Comparison of Signal-to-Noise Ratios for Partial Response Signals With and Without Timing Jitter Immunity

In this section, a comparative study of the Kretzmer[1] signals with and without timing jitter immunity is presented. It is shown that the degradation in signal-to-noise ratio is the same in both cases and the additional bandwidth required for timing jitter suppression guarantees the same signal-to-noise ratio as the signals which have intersymbol interference immunity only. The comparative data for Kretzmer's signals with and without timing jitter immunity is given in Tables I and II, respectively.

TABLE I. Channels without timing jitter immunity.

Signal Class	Partial Response System	$R(\omega)$	Signal-to-Noise Degradation in dB
Ideal	1	T	0
I	$1 + D_L$	$2T\cos(\frac{\omega T}{2})$	2.1
II	$1 + 2D_L + D_L^2$	$4T\cos^2(\frac{\omega T}{2})$	6.0
III	$2 + D_L - D_L^2$	$2T + T[\cos\omega T - \cos 2\omega T] - jT[\sin\omega T - \sin 2\omega T]$	1.2
IV	$1 - D_L^2$	$2jT\sin\omega T$	2.1
V	$-1 + 2D_L^2 - D_L^4$	$4T\sin^2\omega T$	6.0

D_L - delay operator

TABLE II. Channels with timing jitter immunity.

Signal Class	Partial Response System	$R(\omega)$	Signal-to-Noise Degradation in dB
Ideal	1	$T(1 - \frac{T}{2\pi} \lvert \omega \rvert)$	0
I	$1 + D_L$	$2T(1 - \frac{T}{2\pi} \lvert \omega \rvert)\cos(\frac{\omega T}{2})e^{-j\omega \frac{T}{2}}$	2.1
II	$1 + 2D_L + D_L^2$	$4T(1 - \frac{T}{2\pi} \lvert \omega \rvert)\cos^2(\frac{\omega T}{2})e^{-j\omega T}$	6.0
III	$2 + D_L - D_L^2$	$T(1 - \frac{T}{2\pi} \lvert \omega \rvert)[(2T\cos\omega T - \cos 2\omega T) - j(\sin\omega T - \sin 2\omega T)]$	1.2
IV	$1 - D_L^2$	$2jT(1 - \frac{T}{2\pi} \lvert \omega \rvert)(\sin\omega T)e^{-j\omega T}$	2.1
V	$-1 + 2D_L^2 - D_L^4$	$4T(1 - \frac{T}{2\pi} \lvert \omega \rvert)\sin^2(\frac{\omega T}{2})e^{-2j\omega T}$	6.0

D_L - delay operator

Joint Services Technical Advisory Committee
F44620-74-C-0056

M. Kavehrad and L. Kurz

REFERENCES

1. E.R. Kretzmer, "Generalization of a Technique for Binary Data Communications," IEEE Trans. on Comm. Technology, Vol. COM-14 (February 1966).
2. R.A. Gibby and J.W. Smith, "Some Extensions of Nyquist's Telegraph Theory," Bell Syst. Tech. J. (September 1965).
3. A. Lender, "The Duobinary Technique for High Speed Data Transmission," 1963 IEEE Winter General Meeting (1963).

ON ESTIMATION OF CHANNEL CONDITIONS IN M-ARY SYSTEMS WITH RANDOM SIGNALS AND INTERSYMBOL INTERFERENCE

L. Kurz

This report represents our extension of the channel-condition estimation problem considered previously.[1] An estimator for the probability of error of a channel disturbed by additive noise with random signaling and intersymbol interference and M-ary transmission is introduced and analyzed. The estimator is based on simple intuitive notions and is useful for channels where almost all of the intersymbol interference exists in the filters adjacent to the signal (active) filter (for instance, a FSK system). The statistical analysis for the M-ary system is more complicated than the analysis of the previous report.[1] After a few simplifying assumptions, an expression for the mean of the estimator is obtained which is a function of the mean of the estimator of the dependence, ρ, and the noise and signal p.d.f.'s.

A. System Model

The transmitter sends one of m signals ($S_1, S_2, \ldots, S_m$) with known a priori probabilities $p_1, p_2, \ldots, p_m$, respectively. The receiver is composed of m filters with each filter matched to one of the signals. At the end of a signaling interval, the signal classifier chooses which signal had been transmitted on the basis of which filter had the largest sample output. In this model it is known that the outputs from the signal and idle filters are not independent. The exact model to be assumed for the nature of dependence is not known. A typical modulation scheme which corresponds to this model is the M-ary FSK system. In this scheme each of the signals S_i corresponds to a pulse which is centered at frequency f_i. Each of the frequencies, f_i, is different but they are usually spaced closely together to spread the signal energy over as small as possible a range of frequencies. The filters adjacent to the signal filter contain a considerable amount of energy from the signal filter because the pulse shape may not be perfectly shaped, the channel may have some doppler effect on the transmitted frequency and the filters at the receiver are not perfectly orthogonal to the transmitted waveform. In the system model it is then assumed that the intersymbol interference decreases as the distance between the idle filter and the signal filter increases. The estimator of channel conditions for this model is complicated; it requires an estimate of dependence between the active (signal) and idle filters depending on the distance between the active and idle filters. The estimator of the channel conditions would then remove the dependence (to as large a degree as possible) and calculate an estimator of the error probability as if signal and idle filters were independent.

In the proposed model, the signal filter output is composed of two components: a signal which is a sample function of a stochastic process plus additive independent from sample-to-sample noise. Each of the filters adjacent to the signal filter contains an output composed of a function g(S) plus additive noise. The outputs of non-adjacent filters contain noise only. The p.d.f. of the noise is the same for all filters, the function g(S) is the same independent of which signal is transmitted, and the p.d.f. of the output signal is also independent of which signal is transmitted.

If S_i is transmitted, the output of the filter matched to S_i will be denoted by $S_i + N_i$, the adjacent idle filter outputs by $g(S_i) + N_{i-1}$ and $g(S_i) + N_{i+1}$, respectively, and the outputs of the other idle filters by N_j, $j = 1, \ldots, i-2, i+2, \ldots, m$.

B. The True Average Probability of Error

The expression for the probability of error must be broken up into two terms: one for the case where S_1 and S_m are transmitted, and one for the case where S_i is transmitted for $i = 2, \ldots, m-1$. The probability of making an error given S_i is sent is equal to the probability that the signal is equal to the value x_2 multiplied by the probability that the maximum of the idle filter will be greater than x_2 given that the signal filter output is x_2, summed over all values of x_2. The p.d.f. of the maximum of the idle filters will vary depending on which signal is sent. In mathematical terms, we obtain

$$P_e = \sum_{i=1}^{m} P_i P_{e_i}(S_i) = (p_1 + p_m) P_{e_1}(S_1) + \left[\sum_{i=2}^{m-1} p_i\right] P_{e_2}(S_2) \tag{1}$$

where $P_{e_i}(S_i)$ = probability of making an error given that S_i was transmitted and

$$P_{e_1}(S_1) = \int_{-\infty}^{\infty} \int_{x_2}^{\infty} f_{sig}(x_2)\, f_{m_1 n}(x_1, x_2)\, dx_1 dx_2 \tag{2}$$

$$P_{e_2}(S_2) = \int_{-\infty}^{\infty} \int_{x_2}^{\infty} f_{sig}(x_2)\, f_{m_2 n}(x_2, x_2)\, dx_1 dx_2 \tag{3}$$

where

$f_{sig}(x_2)$ = p.d.f. of $S_i + N_i$

$f_{m_1 n}(x_1, x_2)$ = joint p.d.f. of max [m-2 independent noise samples, a sample $g(S_1) + N_2$ given that $S_1 + N_1 = x_2$]

and

$$f_{m_2 n}(x_1, x_2) = \text{joint p.d.f. of max} \begin{bmatrix} \text{m-3 independent noise samples, a} \\ \text{sample } g(S_2) + N_3, \text{ or sample} \\ g(S_2) + N_1 \text{ given that } S_2 + N_2 = x_2 \end{bmatrix}$$

C. Development of the Estimator

The p.d.f.'s, $f_{m_1 n}(x_1, x_2)$ and $f_{m_2 n}(x_1, x_2)$, are not dependent on all the factors that compose the output of the signal filter but are dependent on the part caused by the signal transmitted. In terms of c.d.f.'s

$$F_{m_1 n}(x)\big|_{s_i} = F_n^{m-2}(x)\, F_n(x-g(s_i)) \tag{4}$$

$$F_{m_2 n}(x)\big|_{s_i} = F_n^{m-3}(x)\, F_n^2(x-g(s_i)) \tag{5}$$

The appropriate p.d.f.'s are obtained by differentiating the above equations yielding

$$f_{m_1 n}(x)\big|_{s_i} = (m-2)\, F_n^{m-3}(x)\, F_n(x-g(s_i))\, f_n(x) + F_n^{m-2}(x)\, F_n(x-g(s_i)) \tag{6}$$

$$f_{m_2 n}(x)\big|_{s_i} = (m-3)\, F_n^{m-4}\, F_n^2(x-g(s_i))\, f_n(x) + 2F_n^{m-3}(x)\, F_n(x-g(s_i))\, f_n(x-g(s_i)) \tag{7}$$

and

$$f_{sig}(x_2)\big|_{s_i} = f_n(x_2 - s_i) \tag{8}$$

From Eqs. (4) - (8), it is seen that given the function $g(s_i)$, the p.d.f. of the signal S_i and the p.d.f. of the noise $f_n(x)$, the probability of error can be calculated.

It should be noted that for every value of the signal s_i, the distribution of the idle filters is different. Therefore, the samples from the noise filter taken at one time cannot be compared with the samples taken at another instant of time. If the dependence between the signal and noise filter were removed, all the noise samples can be used to form P_{e_m} - an estimator of P_e. In forming P_{e_m}, an estimation of the linear dependence, ρ, between the signal and idle filter is made. Denoting by A_i the ith output of the active filter, it will be assumed that $A_1 \rightarrow A_p$ are from the first filter, $A_p \rightarrow A_r$ are from the mth filter and $A_{r+1} \rightarrow A_k$ come from the other filters.

The procedure used for approximating $f_{m_1 n}$ and $f_{m_2 n}$ is as follows. All of the samples of the nonadjacent idle filters are received and stored. To approximate $f_{m_1 n}$, given that the output of the active filter is A_i, the value of the nonadjacent idle

filters corresponding to the time A_i was received are first removed. The $f_{m_1 n}$ is then approximated by taking every group of m-1 samples, adding $\rho\, s_{est\ i}$ to the (m-1)th sample and then taking the maximum of the m-1 samples, where $s_{est\ i} = A_i - \overline{N}$ and $\overline{N}$ is the statistical average of all nonadjacent idle filters. The kth of these samples will be denoted by ${}_1M_{i_k}$. The set of samples $\{{}_1M_{i_k}\}$ are an approximation to a set of samples of the distribution with p. d. f. $f_{m_1 n}$ given that the output due to the signal is $s_{est\ i}$. A similar procedure is followed to approximate $f_{m_2 n}$. The estimate of the probability of error contains two terms: one which estimates the probability of error given the first or the last filter is active and one which estimates the probability of error given one of the filters is active, or

$$P_{e_m} = (p_1 + p_m)\frac{1}{rd}\left[\sum_{j=1}^{r}\sum_{i=1}^{d} u(A_j - {}_1M_{j_i})\right] + \left(\sum_{j=2}^{m-1} p_j\right)\left(\frac{1}{k-r}\right)\left(\frac{1}{d}\right)\left[\sum_{i=r+1}^{k}\sum_{j=1}^{d} u\,(A_i - {}_2M_{i_j})\right] \tag{9}$$

where u(.) is the unit step function, and d is the maximum integer of $\frac{(m-2)r+(m-3)(k-r)}{m-1}$. Following a similar procedure to the one outlined in Ref. 1, it can be shown that the estimator P_{e_m} is related to the Mann-Whitney statistic.

D. Estimation of the Linear Dependence Coefficient ρ

In general, g(s) is unknown and it is impossible to develop general procedures to estimate it. If the dependence is weak, as is the case in most practical systems, satisfactory answers are obtained if g(s) is replaced by ρ s. The linear coefficient, ρ, can then be estimated using

$$\rho = \frac{1}{2k-r}\left[\sum_{j=1}^{p} A_j I_{a_{j_u}} + \sum_{j=p+1}^{r} I_{a_{j_e}} A_j + \sum_{j=r+1}^{k} (I_{a_{j_u}} + I_{a_{j_e}})\, A_j\right] - \overline{I}\,\overline{A}\,\sigma^2_{s_i}$$

where

A_j = output of the active filter at the jth instant

$I_{a_{ju}}$ = output of the adjacent idle filter at the jth instant which corresponds to the filter which is matched to a signal greater than the active filter

$I_{a_{je}}$ = output of the adjacent idle filter at the jth instant which corresponds to the filter which is matched to a signal less than the active filter

$$\overline{I} = \frac{1}{2k-r}\left[\sum_{j=1}^{p} I_{a_{ju}} + \sum_{j=p+1}^{r} I_{a_{je}} + \sum_{j=r+1}^{k} (I_{a_{ju}} + I_{a_{je}})\right]$$

$$\overline{A} = \frac{1}{k} \sum_{j=1}^{k} A_j$$

and $\sigma^2_{s_i}$ = variance of the signal

In deriving the expression for the estimator of P_e the assumption was made that the decisions at the receiver were correct. This assumption affects the accuracy of the estimator insignificantly.

Joint Services Technical Advisory Committee
F44620-74-C-0056

L. Kurz

REFERENCES

1. L. Kurz, "On Estimation of Channel Conditions in Digital Communication Systems," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-75, pp. 268-274 (November 1976).

A NEW CLASS OF NON-RANK ROBUST DETECTORS

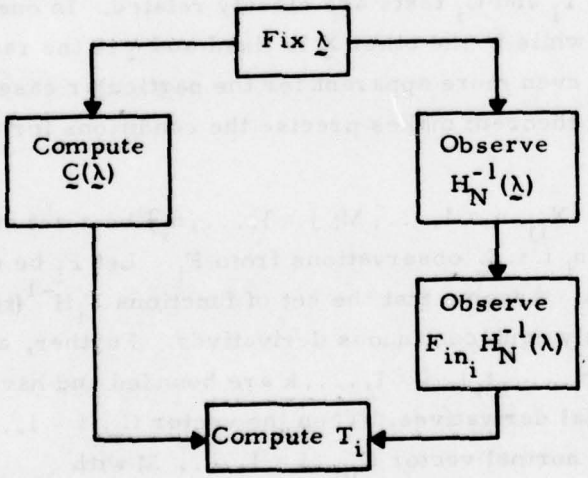
J. I. Cochrane and L. Kurz

In this report, a new class of non-rank robust detectors, which are related to the generalized quantile detector,[1] is considered. Unlike the non-rank detectors based on quantile statistics with fixed scoring vectors, the new class of detectors uses a random scoring vector. The new statistics, called C_i statistics, includes the Ching-Kurz statistics[2] as a special case and allow useful asymptotic expressions for a wide class of linear non-rank tests. The notation used in this report is the same as in Reference 1.

A. Asymptotic Normality of the C_i Statistics

The T_i statistics of Ref. 1 are designed in three interrelated steps. First, the vector $[\beta_1, \ldots, \beta_M]$ is chosen to achieve the desired weighing of the empirical distribution functions. Second, the vector $[\lambda_1, \ldots, \lambda_m]$ is chosen to maximize some figure of merit. Third, the best scores $[C_1, \ldots, C_m]$ for that $\underset{\sim}{\lambda}_{opt}$ are chosen to maximize some, possibly different, figures of merit. For example, the $\underset{\sim}{\lambda}_{opt}$ may be selected to guarantee robustness over a wide class of distributions, while $\underset{\sim}{C}_{opt}$ is chosen to maximize the efficacy for a particular distribution.

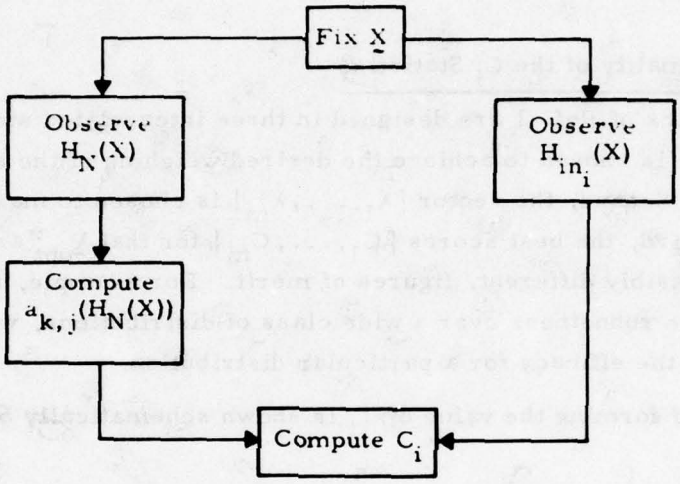
The process of forming the value of T_i is shown schematically below:



In this section we form the new C_i statistics by using these same steps in a different order to avoid the ranking necessary to observe H_N^{-1}. Let

$$C_i = \sum_{j=1}^{k} a_{k,j}(H_N(X_1), \ldots H_N(X_k))F_{in_i}(X_j) \qquad (1)$$

where $[X_1 < \ldots < X_k]$ is an arbitrary set of thresholds and $a_{k,j}(t_1, \ldots, t_k)$, $j = 1, \ldots, k$ is a set of scoring functions defined on $[0 < t_1 \ldots t_k < 1]$. These statistics are formed by the process shown as follows



When $a_{k,j} = C_j$, $j = 1, \ldots, k$, Eq. (1) reduces to the m-interval tests of Ching and Kurz.[2] Intuitively, the T_i and C_i tests are closely related. In one case $\underset{\sim}{\lambda}$ is fixed and $\underset{\sim}{X}$ is a random vector, while in the other $\underset{\sim}{X}$ is fixed and $\underset{\sim}{\lambda}$ is the random vector. The similarity will become even more apparent for the particular cases considered in this section. The following theorem makes precise the conditions for asymptotic normality of the C_i statistics.

Theorem 1. Let $\{X_{ij}, i = 1, \ldots, M; j = 1, \ldots, n_i\}$ be a set of observations, where X_{ij} is the set of n_i i.i.d. observations from F_i. Let F_i be a set of strictly monotonic distributions. Assume that the set of functions $F_i H^{-1}(t)$ on $[\epsilon, 1-\epsilon]$ for any $\epsilon > 0$ have uniformly equi-continuous derivatives. Further, assume that the scoring functions $a_{k,j}(t_1, \ldots, t_k)$, $j = 1, \ldots, k$ are bounded and have uniformly equi-continuous second partial derivatives. Then the vector C_i, $i = 1, \ldots, M$ converges in distribution to a jointly normal vector C_{io}, $i = 1, \ldots, M$ with

$$E\{C_{io}\} = \sum_{j=1}^{k} a_{k,j}(\lambda_1, \ldots, \lambda_k)F_i H^{-1}(\lambda_j) \qquad (2)$$

where $[\lambda_1, \ldots, \lambda_k]$ is a set of parameters uniquely defined by $\lambda_i = H(X_i)$, and

$$\operatorname{cov}\{C_{io}, C_{jo}\} = \sum_{\mu=1}^{M} n_\mu^{-1} [\delta_{iu} \underset{\sim}{a} + \beta_\mu \underset{\sim}{C}_\mu] \Sigma_\mu [\delta_{j\mu} \underset{\sim}{a} + \beta_\mu \underset{\sim}{C}_\mu] \tag{3}$$

where δ_{ij} is the Kronecker delta and

$$\underset{\sim}{a}^T = [a_{k,1}, \ldots, a_{k,k}]$$

$$C_u^T = [C_{iu}, \ldots, C_{ku}]$$

$$C_{iu} = \sum_{j=1}^{k} \left. \frac{\partial a_{k,j}}{\partial t_u} \right|_{t_u = \lambda_u} F_u H^{-1}(\lambda_j) \tag{4}$$

and $\sum_u$ is a k by k matrix with elements

$$\sigma_{iju} \quad F_u H^{-1}(\lambda_i)[1 - F_u H^{-1}(\lambda_j)] \quad i \leq j \tag{5}$$

Proof. (Because of its length and complexity, the proof is omitted.) Under the null hypothesis, the variance simplifies sharply, since $C_{iu} = C_{iv}$ and $\sum_u = \sum_v$ for $u, v = 1, \ldots, M$.

For a particular choice of $\underset{\sim}{X}$, it is possible to choose the score functions and their derivatives to control the asymptotic distribution of C_i. In sections B and C two design techniques for the class of C_i statistics are suggested.

B. Nonparametric Aspects of the C_i Statistics

The distribution of C_i is dependent on the quantiles of the underlying distribution, even under the null hypothesis, and the nonparametric property of fixed Type I error for the tests of Ref. 1 is lost. If the only problems of interest were binary tests with the Neyman-Pearson formulation (constant-size test), this is a serious drawback. If, however, the statistics are designed for binary tests with minimum probability of error or for more general composite hypothesis testing, the nonparametric property is less important than other aspects of the desired robustness. This point may best be illustrated by a simple example.

Consider a two-sample problem using the Mann-Whitney statistic. In asymptotic analysis, the separation between the expected value of the test statistics is

$$E\{T_1 | H_1\} - E\{T_1 | H_o\} \propto \int_{\infty}^{\infty} F_1 dF_2$$

and if minimum probability of error is to be achieved (asymptotically) the decision must be to accept H_o if $T_1 \leq C$ and H_1 if $T_1 \geq C$ where

$$C \propto \int_{-\infty}^{\infty} F_1 dF_2$$

But the choice of this threshold cannot be made without a priori knowledge of this parameter or an adaptive or learning procedure to set this threshold. In this example, the nonparametric property of the rank tests is not helpful in selecting the decision partition. In M-ary and composite hypothesis testing problems, the choice of critical regions in the decision space spanned by the intermediate statistic is always dependent on some parameters of the underlying distribution. In this sense, the absence of the nonparametric property of fixed Type I error in non-rank tests is not important.

The asymptotic distribution of C_i is expressed in terms of $\underset{\sim}{\lambda}$, which is not known for a general nonparametric problem. This difficulty may be approached in two ways. First, in stationary noise, it is possible to use iterative techniques to form estimates of λ for a fixed X which converge with probability one to the true value of H(X). Alternately, $\underset{\sim}{X}$ may be controlled by stochastic approximation techniques so that they converge with probability one to the desired values $H^{-1}(\lambda)$ for a fixed set $\underset{\sim}{\lambda}$. Second, it is possible to choose scoring functions so that the expected value and variance of the C_i statistics are locally distribution-free. Using this techniques, the C_i statistics are shown in Section D to be asymptotically locally equivalent to the T_i statistics of Reference 1.

C. Minimum-Variance Scores

A useful figure of merit for asymptotic analysis of the C_i statistics is $Tr[\Gamma_M]$ where Γ_M is the MxM covariance matrix of C_i, i = 1, . . . , M. For normal vector statistics, this is the average variance of the independent set of random variables $\underset{\sim}{B} = T\underset{\sim}{C}$ where $T^T \Gamma T = \text{diag}[\sigma_1^2, \ldots, \sigma_M^2]$. For the normal C_{io} specified by Eqs. (2) and (3) under the null hypothesis.

$$Tr\{\Gamma_o\} = \left[\sum_1^M n_i^{-1}\right] \underset{\sim}{a}^T \sum \underset{\sim}{a} + 2 \left[\sum_1^M \beta_i n_i^{-1}\right] \underset{\sim}{C}^T \sum \underset{\sim}{a} + M \left[\sum_1^M \beta_i^2 n_i^{-1}\right] \underset{\sim}{C}^T \sum \underset{\sim}{C} \qquad (6)$$

where $\underset{\sim}{C}$, $\underset{\sim}{a}$ and $\sum$ are defined in Section A. Since $\underset{\sim}{C}$ does not appear in the expected value of $C_{i,o}$, it is possible to select $\underset{\sim}{C}$ to minimize $Tr\{\Gamma_o\}$ independent of $\underset{\sim}{a}$. This minimization yields

$$\underset{\sim}{C}_{opt} = \frac{-\sum_{1}^{M} \beta_i n_i^{-1}}{M \sum \beta_i^2 n_i^{-1}} \underset{\sim}{a} \tag{7}$$

and

$$\min_{C} \{Tr[\Gamma_o]\} = \underset{\sim}{a}^T \sum \underset{\sim}{a} \left[\sum n_i^{-1} - \frac{[\sum \beta_i n_i^{-1}]^2}{M \sum \beta_i^2 n_i^{-1}} \right] \tag{8}$$

The selection of the $\{\beta_i\}$ and $\underset{\sim}{a}$ vectors for a particular problem must consider more complex figures of merit; for $\underset{\sim}{C}_{opt}$ given by $\underset{\sim}{C} = k\underset{\sim}{a}$ and a given $\{\beta_i\}$, the asymptotic covariance of C_i under the null hypothesis is identical to within a constant to the T_i statistics of Reference 1. The asymptotic expected values are identical. All of the techniques and results of Ref. 1 are thus applicable to the C_i design problems.

It is interesting to note that Eq. (7) is minimized over $\underset{\sim}{\beta}$ by selecting $\beta_1 = \ldots = \beta_m = 1/m$ and the resulting variance is given by

$$\min_{\underset{\sim}{\beta}} \min_{\underset{\sim}{C}} Tr\{\Gamma_o\} = \underset{\sim}{a}^T \sum \underset{\sim}{a} \left(\frac{M-1}{M}\right) \left(\sum n_i^{-1}\right) \tag{9}$$

This rather surprising result is verified by noting that $\beta_1 = \ldots = \beta_m$ achieves equality for the Stieltjes integral formulation of the Cauchy-Schwartz inequality. For that special case

$$\underset{\sim}{C}_{OPT} = -\underset{\sim}{a} \tag{10}$$

D. Locally T_i Equivalent C_i Statistics

In the previous discussions, it has been emphasized that threshold statistics are not nonparametric under the null hypothesis. In particular, the expected value of C_i is a function of the unknown values $\lambda_i = H(X_i)$. In this section, we seek to minimize this uncertainty in the expected value under the null hypothesis by suitable choice of the scoring function. From the definition of $\underset{\sim}{C}$ in Eq. (3), the only influence that can be applied to $a_j(\lambda_1, \ldots, \lambda_k)$ to control the asymptotic expected value is the set of first partial derivatives of a_j. We choose those partials so that the expected value remains approximately constant in a neighborhood of the design values $\lambda_1, \ldots, \lambda_k$, by setting

$$\sum_{1}^{k} \left(a_i + \Delta\lambda_j \frac{\partial a_i}{\partial \lambda_j}\right) (\lambda_i + \delta_{ij} \Delta \lambda_j) = \sum_{1}^{k} a_i \lambda_i \tag{11}$$

which is equivalent to setting

$$\underset{\sim}{C} = -\underset{\sim}{a} \tag{12}$$

where the terms of order Δ^2 have been neglected. Using this value insures that the expected value of C_i will remain constant to within a term of order $n_{min}^{-1/2}$ for small errors in setting the thresholds $[X_1, \ldots, X_m]$. From the previous section, this is also the solution for $\underset{\sim}{C}$ which minimizes the average variance when $\beta_1 = \ldots \beta_m$.

Joint Services Technical Advisory Committee
F44620-74-C-0056

J.I. Cochrane and L. Kurz

Bell Telephone Laboratories

REFERENCES

1. J.I. Cochrane and L. Kruz, "A Generalized Quantile Detector," Progress Report No. 41 to JSTAC, Polytechnic Institute of New York, Report No. R-452.41-76, pp. 315-321 (November 1976).

2. Y.C. Ching and L. Kurz, "Nonparametric Detectors Based on m-Interval Partitioning," IEEE Trans. on Information Theory, Vol. IT 18, No. 2, pp. 251-257 (March 1972).

THE M-INTERVAL PARTITION DETECTOR (MIPD) WITH DEPENDENT INPUT DATA

P. Kersten and L. Kurz

Detectors that are nonparametric in the sense that the Type I error is independent of the underlying distribution, lose this property once the samples used become dependent. In the past it has been assumed that the samples used in the MIPD are independent. Once this assumption is dropped, the performance of this test under the hypothesis cannot be guaranteed. However, the situation can be circumvented to some extent. Woinsky and Kurz[1] proposed a class of detectors which retain their nonparametric character with dependent samples. By properly grouping the dependent samples one creates a sequence of independent vector samples of dimension n. This is accomplished by placing n consecutive samples in a vector of length n and then skipping $\gamma-1$ samples to ensure that the next vector of n components is independent of the first vector. This assumes that one is sampling at γ times the rate required to obtain independent samples $\gamma = 1, 2, \cdots$. One then applies a fixed transformation from n-dimensional to one-dimensional space yielding a sequence of independent random variables. Mathematically, this can be stated using the notation of Ref. 2 in the following manner. Let $X_1, \cdots, X_N$ be a sequence of independent samples with a total time of observation of NT. The sampling rate is then increased by a factor γ and the resulting sequence is denoted by $Y_1, \cdots, Y_{N\gamma}$. Grouping n consecutive Y samples into a vector $\underline{Z}$ the resulting sequence of vectors is denoted by $\underline{Z}_1, \cdots, \underline{Z}_{[N/(n+\gamma-1)]}$ where $[x \cdot y] = x$ and $\underline{Z}_i = (Y_{(Nn+\gamma-1)(i-1)+1}, \cdots, Y_{(n+\gamma-1)(i-1)+n})$.

Under the null hypothesis the $\underline{Z}_i$ have a joint distribution function given by $F(x_1, \cdots, x_n)$ and under the alternative they are samples from a noise distribution which includes the signal. After passing through the predetector which is actually just a transformation $\mathcal{L}$, the resulting sequence $\mathcal{L}(Z_1), \cdots, \mathcal{L}(Z_L)$, where $L = [N\gamma/(n+\gamma-1)]$, has a distribution F_z under the null hypothesis and G_z under the alternative. The samples are used in a standard nonparametric detector to obtain a constant false alarm rate. The central idea which makes this model so simple and yet so effective is to concentrate the effect of the dependence within the transformation and thus allow the detector to operate in a nonparametric mode.

The effect of the transformation is to create a new sample space whose underlying distribution is determined by $\mathcal{L}$ and the distribution of the samples Y. The transformation $\mathcal{L}$ is selected to be the sample mean.[1] Note, that for independent Gaussian samples with known variance this is a sufficient statistic.

Following Ref. 3, one evaluates the performance of this detector against another detector via the asymptotic relative processing time ARPT which is defined to be

$$ARPT(T_1, T_2) = ARE(T_1, T_2) \times (t_2/t_1)$$

where the $ARE(T_1, T_2)$ is the asymptotic relative efficiency of the first detector with respect to the second and t_i is the time between samples for detector T_i. Note that if $t_2 = t_1$, this implies that all the samples are independent and the ARPT reduces to the ARE.

For many symmetric distributions of T_i, the transformation

$$\mathcal{L}(Y_1, \cdots, Y_n) = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

will be approximately normally distributed with zero mean and with variance

$$\sigma^2 = \sigma_Y^2 \left(1 + 2 \sum_{k=1}^{\min(\gamma_1 n)-1} (1 - \frac{k}{n}) \rho_Y(kT/\gamma)\right) \tag{1}$$

where ρ_Y is the correlation coefficient of Y_i and $Y_{i+KT/\gamma}$. In general, define

$$Z^i = \frac{1}{n} \sum_{j=1}^{n} Y_{(n+\gamma-1)(i-1)+j}.$$

To apply this to the m-Interval Partition Test, one now uses the samples Z^i to obtain the quantiles of F_z. The MIPD and its first moment are given by

$$T_z = \frac{1}{L} \sum_{j=1}^{L} \sum_{i=1}^{m} \beta_i I_{\{Z^j \varepsilon a_i\}}$$

and

$$E(T_z \mid H_o) = \frac{1}{L} \sum_{j=1}^{L} \sum_{i=1}^{m} \beta_i P(Z^j \varepsilon a_i) = \frac{1}{m} \sum_{i=1}^{m} \beta_i = 0$$

where

$$a_i = (\alpha_{i-1}, \alpha_i], \quad \alpha_i = F_z^{-1}(i/m).$$

Accordingly,

$$Var(T_z \mid H_o) = \frac{1}{mn} \sum_{i=1}^{m} \beta_i^2$$

and

$$E(T_z \mid H_1) = \sum_{i=1}^{m} \beta_i (F_z(\alpha_i - \theta) - F_z(\alpha_{i-1} - \theta)).$$

To obtain the efficacy one evaluates

$$\partial E(T_z|H_1)/\partial\theta = \sum_{j=1}^{m} \beta_i[f_z(\alpha_{i-1}) - f_z(\alpha_i)]$$

and thus

$$\mathcal{E}(T_z) = m(\sum_{i=1}^{m} \beta_i[f_z(\alpha_{i-1}) - f_z(\alpha_i)])^2/\sum_{i=1}^{m} \beta_i^2 .$$

On the other hand, for the MIPD constructed upon the independent samples X_i one obtains an efficacy of the exact same form with β replaced with b, α replaced by a and f_z replaced with f_x. Therefore, the asymptotic relative processing time is

$$ARPT(T_z, T_x) = (\gamma/(n+\gamma-1)(\mathcal{E}(T_z)/\mathcal{E}(T_x)).$$

To examine a specific case, it is assumed that both T_z and T_x use the locally most powerful scores,[4] i.e., $\beta_i = f_z(\alpha_{i-1}) - f_z(\alpha_i)$ and b_i by the same formula with α replaced with a and f_z replaced with f_x. Then the

$$ARPT(T_z, T_x) = (\gamma/(\gamma+n-1))(\sum_{i=1}^{m} \beta_i^2/\sum_{i=1}^{m} b_i^2).$$

If $\gamma = 1$ so that $Y_i = X_i$ are i.i.d. N(0, 1) samples, then $P(Z \leq x) = \Phi(\sqrt{n}\, x)$ and $\Phi(\sqrt{n}\, \alpha_i) = i/m = \Phi(a_i)$ which implies $\alpha_i = a_i/n^{1/2}$. Then

$$f_z(\alpha_i) = (n/2\pi)^{1/2} \exp(-\alpha_i^2 n/2) = (n/2\pi)^{1/2} \exp(-a_i^2/2) = n^{1/2} f_x(a_i)$$

so $\beta_i^2 = nb_i^2$ and the ARPT = 1. However, should the samples be dependent, the ARPT ARPT = $(\gamma/(n+\gamma-1))/\sigma^2$ where σ^2 is the variance given by Equation (1). This is Eq. (6) of Ref. 3, and shows that the ARPT increases as shown in Figs. 1-3 of the same reference, precisely as the sign statistic and Wilcoxon Test Statistic.

The performance of both the Wilcoxon and Sign Test when the noise is from a Double-Exponential distribution is degraded with respect to its performance in a Gaussian noise environment. The ARPT for the Double-Exponential is $2/\pi$ times and $1/\pi$ times the ARPT for the Gaussian CDF for the Wilcoxon and Sign Test respectively. One suspects that the ARPT for the MIPD should be greater than that of the Sign Test under the same comparison. Calculation reveals that for m = 4 partitions the corresponding multiplicative factor is 0.43. Although this is not as good as the Wilcoxon Statistic, one must remember that the MIPD is considerably easier to implement than the Wilcoxon Test. For m > 4 the gap between the two tests is essentially closed.

Joint Services Technical Advisory Committee
F44620-74-C-0056

P. Kersten and L. Kurz

REFERENCES

1. M.N. Woinsky and L. Kurz, "Nonparametric Detection Using Dependent Samples," IEEE Trans. on Information Theory, pp. 355-358 (May 1970).

2. S.A. Kassam and J.B. Thomas, "A Class of Nonparametric Detectors for Dependent Input Data," IEEE Trans. on Information Theory, Vol. IT-21, No. 4, pp. 431-437 (1975).

3. A.P. Sage, "Optimum Systems Control," Englewood Cliffs, N.J.: (Prentice Hall, 1968).

4. Y.C. Ching and L. Kurz, "Nonparametric Detectors Based on m-Interval Partitioning," IEEE Trans. on Information Theory, Vol. IT-18, No. 2, pp. 251-257 (March 1972).

A ROBUSTIZED CUMULATIVE DECISION FEEDBACK SYSTEM

H. S. Ashtiani and L. Kurz

Since it was proved that transmission with vanishingly small errors is possible, at least in theory, as long as the information rate is below channel capacity, there has been a continuous effort on the part of investigators to devise methods to achieve or at least approach this goal.

This effort has been pursued in several directions. One direction has been to develop sophisticated coding and decoding schemes to detect and correct errors occurring during transmission so that the final probability of error can be kept arbitrarily low; this method, in many cases, has resulted in long sequences of coded messages and complex coding and decoding schemes that are not very desirable from a practical point of view.

Another approach has been to include a feedback link in the transmission system. This approach takes into account the fact that in practice a second link is often available whether or not it is used as a feedback link.

The price that has to be paid for improving the error rates, assuming fixed signal power, in both cases is a lower information rate. Attention has been paid to find ways of improving error rates without compromising the information rate of the system too much. In the first approach this has resulted in more efficient coding techniques, and in the second approach in feedback systems that do not disregard the ambiguous messages, and use more of the information received, e.g., cumulative or sequential binary decision feedback system (CBDFS).

In this report, we attempt to robustize a sequential detector known as the cumulative binary decision feedback system (CBDFS).[1]

By robustizing, in the context of the decision feedback system under consideration, we mean desensitizing the performance of the system to impulsive noise. This is accomplished by means of setting thresholds in the decision device, and identifying the samples falling in the regions beyond the thresholds as ambiguous samples. In the cumulative system under consideration these samples are not disregarded; however, the decision is postponed until the sum of the repeated sequence of samples lie in regions identified as "safe" regions for making decision. These safe regions lie between a null zone[2] and the two regions beyond impulsive noise thresholds. The null zone is included to protect against decisions made on small samples considered ambiguous as the result of Gaussian noise interference.

A. The Cumulative Binary Decision Feedback System in Gaussian Noise

The CBDFS in Gaussian noise has been discussed and formulated in Reference 1. In this system the receiver integrates the sequence of repeats of a binary digit (the message unit) until the integrated signal passes the null zone boundaries $-k_1$ or k_1, in which case "zero" or "one" is accepted.

The detection problem of finding the average detection time (or average sample number) corresponds to the first-passage time in the random walk problem.[3]

Let $P_j(z, r)_{,\ j=0,1}$ represent the joint probability, when x_j is being sent, that after r transmission, the integrated signal is z, and that z has remained ambiguous (i. e., in the null zone) on all previous transmissions. In the limiting case, when the number of transmissions per message digit is large, the central limit theorem is applicable, so that the integrated signal has a Gaussian distribution in the absence of the null zone boundaries, $P_j(z, r)$ may be evaluated, for large r, by evaluating the corresponding quantity $P_j(z, t)$. The integration time is given by $t = r/a$, a being the number of transmissions per unit time.[1]

From the random walk problem $P_j(z, t)$ is given as a solution of the partial differential equation

$$\frac{\partial P_j(z,t)}{\partial t} = B_j \frac{\partial P_j(z,t)}{\partial z} + D_j \frac{\partial^2 P_j(z,t)}{\partial z^2}$$

subject to the boundary conditions

$$P_j(k_1, t) = P_j(-k_1, t) = 0$$

$$P_j(z, o) = \delta(z)$$

and where

$$B_j = -as_j$$

$$D_j = a\,N/2$$

with s_j, the mean of the received signal, and N, the corresponding variance.

The average probability of error and the ASN for a symmetrical system with equal a priori digit probability and when the digits are transmitted by sending s through a channel of additive Gaussian noise of power N is given in Reference 1.

$$P_e = \frac{1}{1 + e^{2\,s/N\,k_1}}$$

$$\bar{n} = \frac{k_1}{s}\,\frac{e^{2\,s/N\,k_1} - 1}{e^{2\,s/N\,k_1} + 1}$$

B. The CBDFS in Impulsive Noise

In this section the performance of the cumulative binary decision feedback system is analyzed in an environment of impulsive noise.

The impulsive noise is often identified by a distribution having a large variance. This characteristic is reflected in the tail of the distribution (i. e., thicker tails). One problem associated with analyzing the effects of impulsive noise in communication systems is finding a suitable distribution that has the thick tails indicated by an impulsive noise interference and that, at the same time, permits mathematically tractable analysis. One such model is given by functions of the form

$$f(x) = Ke^{-\gamma|x|^{\eta}}, \quad 1 < \eta < 2 .$$

A simple yet reasonably accurate characterization is obtained using the Laplacian distribution $f_L(x) = \gamma/2\, e^{-\gamma|x|}$. In this section the CBDFS is analyzed in an impulsive noise environment, using this distribution.* With a Laplacian distribution, the Fokker-Plank equation of Section A takes the form

$$\frac{\partial P_j(z,t)}{\partial t} = \mathrm{SGN}\,(z)\,\frac{\partial P_j(z,t)}{\partial z} + 1/\gamma\,\frac{\partial^2 P_j(z,t)}{\partial z^2}, \quad j = 0, 1 \tag{1}$$

where $\mathrm{SGN} = \begin{cases} 1 & z > 0 \\ -1 & z < 0 \end{cases}$

The boundary conditions are the same as in Section A. Assuming symmetrical thresholds k_2 and $-k_2$ these boundary conditions are

$$\begin{aligned} &P_j(k_2,\ t) = P_j(-k_2, t) = 0 \\ &P_j(z, 0) = \delta(z) \end{aligned} \tag{2}$$

Using the classical method of separation of variables, Eq. (1) is solved subject to the boundary conditions Eq. (2), yielding

---

* A more general noise model in this class is treated by R. Chassaing and L. Kurz elsewhere in this report.

$$P_1(z,t) = \frac{1}{k_2}\, e^{-\left[\frac{\gamma t}{4} + \frac{\gamma z\, SGN(z)}{2}\right]} \sum_n e^{-\left(\frac{n^2\pi^2 t}{4\gamma k_2^2}\right)} \cos\frac{n\pi}{2k_2} z$$

Following similar steps as in Ref. 1, assuming a symmetrical system with equal a priori digit probability, i.e., when the binary digits are transmitted by sending constant levels s and -s, through a channel of additive impulsive noise of density function $f_L(x) = \gamma/2e^{-\gamma|x|}$, the average probability of error and the ASN are obtained

$$P_e = \frac{}{1+e^{\gamma k_2}}, \quad \bar{n} = \frac{k_2}{s}\,\frac{e^{\gamma k_2}-1}{e^{\gamma k_2}+1}$$

γ is given in terms of the impulsive noise power, N

$$Var(x) = \frac{2}{\gamma^2} = N$$

C. A Robustized CBDFS

In this section we develop a robustized version of the cumulative binary decision feedback system discussed in Sections A and B.

Consider the case of binary signal transmission through a noisy channel, where noise is assumed to be a mixture of Gaussian and impulsive noise. This mixture is represented by a c.d.f. of the form

$$F(x) = (1-\epsilon)\, P(x) + \epsilon\, H(x)$$

where P(x) is a low variance Gaussian distribution representing the background noise, and H(x) is a large variance distribution representing the occasional excursions of impulsive noise.

The purpose of this section is to improve the reliability of the performance of the CBDFS in impulsive noise through desensitizing its performance to the impulsive noise component of the mixed distributed interference.

As suggested by Rappaport and Kurz,[4] this improvement may be obtained by the addition of an upper and a lower threshold, k_2 and $-k_2$ in the detector. As the result, the signal region is divided into five regions. Two of these regions are regions of acceptance of the two binary levels; the upper is the region of acceptance of "1," and the lower is the region of acceptance of "-1." The middle region is the null zone or the region of small signal ambiguity, and the two outer regions are regions of large signal ambiguity (see Figure 1).

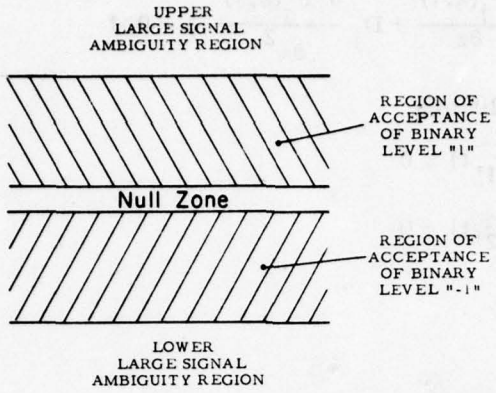


Fig. 1. Decision thresholds.

During transmission, the transmitter repeats a digit, and the receiver integrates the received sequence of repeats. When the integrated signal, z, crosses one of the four boundaries $-k_2$, $-k_1$, k_1 or k_2, the receiver accepts "1" if $k_1 < z < k_2$ and accepts "-1" if $-k_2 < z < -k_1$. The transmitter is then asked, via the feedback link, to start sending the next digit.

D. RCBDFS Formulation

The analysis of the CBDFS is considerably simplified by the assumption that the impulsive noise component of the mixture distribution is of bounded variance.

This assumption permits us to invoke the central limit theorem, and as the result, the relatively simple form of the Fokker-Plank equation remains valid. However, the boundary conditions are modified by the introduction of an additional set of thresholds in the received signal space.

Let $P_j(z, r)_{j=0,1}$ represent the joint probability, when x_j is being sent, that after r transmission the integrated signal is z, and that z has remained ambiguous (i.e., either in the null zone or in one of the two large signal ambiguity regions) on all previous transmissions. In the limiting case, when the number of transmission per message digit is large, the central limit theorem is applicable, so that the integrated signal has a Gaussian distribution in the absence of the ambiguity region boundaries; this Gaussian distribution corresponds to the sum of two Gaussian distributed r.v.'s each corresponding to one of the components of the mixture noise; $P_j(z, r)$ may be evaluated, for large r, by evaluating $P_j(z, t)$, where integration time, t, is given by $t = r/a$, where a is the number of transmissions per unit time.

From the random walk problem $P_j(z, t)$ is given as a solution to

$$\frac{\partial P_j(z,t)}{\partial t} = B_j \frac{\partial P_j(z,t)}{\partial z} + D_j \frac{\partial^2 P_j(z,t)}{\partial z^2} \quad j = 0, 1 \tag{3}$$

subject to the boundary conditions

$$P_j(k_1, t) = P_j(-k_1, t) = 0$$

$$P_j(k_2, t) = P_j(-k_2, t) = 0$$

$$P_j(z, 0) = \delta(z) \tag{4}$$

where $B_j = -as_j$

$D_j = a\,N/2$

With s, the mean of the received signal and N, the variance of the mixture noise distribution (i.e., sum of the two variances of the two Gaussian components). Using the method of separation of variables Eq. (3) is solved subject to the boundary conditions Equation (4)

$$P_j(z,t) = \left(\frac{1}{k_1} + \frac{1}{k_2}\right) e^{\left[-\frac{B_j^2 t}{4D_j} - \frac{B_j}{2D_j} z\right]} \sum_n \left\{ e^{\frac{-n^2\pi^2 D_j t}{4}\left(\frac{1}{k_1^2} + \frac{1}{k_2^2}\right)} \cdot \cos\frac{n\pi}{2k_1} z \cos\frac{n\pi}{2k_2} z \right\}$$

Following similar steps as in Ref. 1, the average probability of error, and the ASN for a symmetrical bipolar binary signal of level s, transmitted through a channel of additive Gaussian and impulsive noise of total power N, are given by

$$P_e = \frac{1}{1 + e^{2\frac{s}{N}\frac{k_1 k_2}{k_1 + k_2}}}$$

$$\bar{n} = \frac{k_1 k_2}{s(k_1 + k_2)} \frac{e^{2\frac{s}{N}\frac{k_1 k_2}{k_1 + k_2}} - 1}{e^{2\frac{s}{N}\frac{k_1 k_2}{k_1 + k_2}} + 1}$$

Figure 2 shows the performance of the CBDFS, with symmetrical null zone boundaries $-k_1$ and k_1, in the mixture noise environment in terms of ASN as a function of signal-to-noise ratio, at a fixed error rate. For simplicity the signal

level is fixed at 1. Two impulsive noise intensities, $\epsilon = 5\%$ and $\epsilon = 10\%$, and two error rates, $P_e = .01$ and $P_e = .0001$, are assumed.

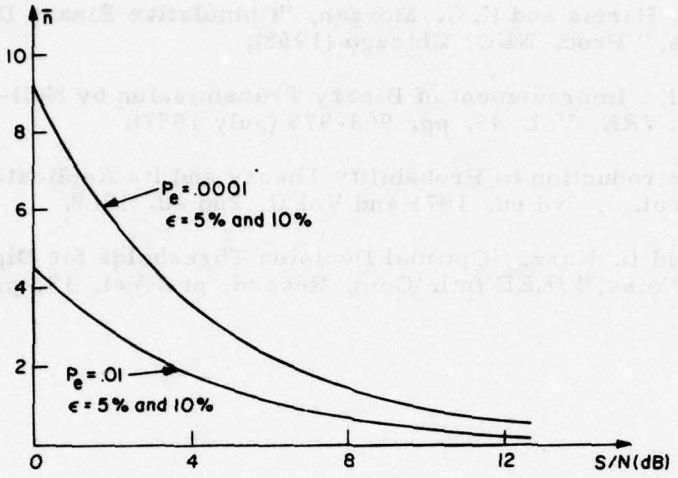


Fig. 2. Performance of CBDFS.

In Fig. 3 the performance of the robustized CBDFS, with symmetrical impulsive noise thresholds $-k_2$ and k_2, in the same mixture noise environment as the CBDFS and under the same assumptions, are given. It is seen that a significant reduction in the average sample size of CBDFS is achieved as the result of protecting its performance against impulsive noise. This improvement was expected because the CBDFS is a parametric detector and is not robust with respect to variations in the noise distribution, e.g., occurrence of impulsive noise.

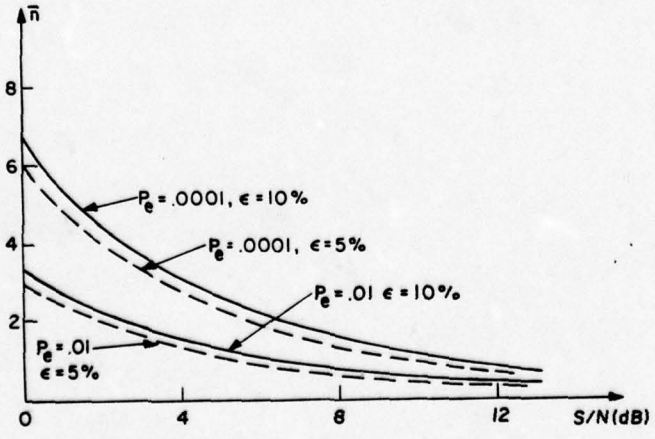


Fig. 3. Performance of robustized CBDFS.

Joint Services Technical Advisory Committee
F44620-74-C-0056

H. S. Ashtiani and L. Kurz

REFERENCES

1. S. S. L. Chang, B. Harris and K. C. Morgan, "Cumulative Binary Decision Feedback Systems," Proc. NEC, Chicago (1958).

2. F. J. Bloom, et al, "Improvement of Binary Transmission by Null-Zone Reception," Proc. IRE, Vol. 45, pp. 963-975 (July 1957).

3. W. Feller, "An Introduction to Probability Theory and Its Application," John Wiley and Sons, Vol. I, 3rd ed. 1971 and Vol II, 2nd ed. 1968.

4. S. S. Rappaport and L. Kurz, "Optimal Decision Thresholds for Digital Signalling in Non-Gaussian Noise," IEEE Intl. Com. Record, pt 2 Vol. 13, pp. 198-212 (1965).

SEQUENTIAL PARTITION DETECTORS WITH DEPENDENT SAMPLING

R.F. Dwyer and L. Kurz

In this report, the theory of sequential partition detectors with dependent sampling is introduced. A new formulation is given which predicts the thresholds under q-dependent sampling in order to maintain the same error probabilities as in the independent sampling case. A comparison is made between independent and dependent sequential partition detectors based on the average time to detection. Under stated conditions dependent sequential partition detectors show improved efficiency for both Lehmann and shift of the mean alternatives.

A. Sequential Partition Detectors (SPD)

It was shown in Ref. 1 that the SPD test statistic was given by

$$T_n = \sum_{i}^{n} \sum_{k}^{m} b_k n_{ik} \quad , \tag{1}$$

where $b_k = \ell n(P_{1k}/P_{ok})$, are called scores and n_{ik} is a counting device for keeping track of which interval the observations fall in.

Thus, Eq. (1) represents the classical cumulative sum form for the SPD with stopping boundaries given by

$$b = \ell n B < T_n < \ell n A = a \quad ,$$

where the test is terminated if $T_n \geq a$ or $T_n \leq b$ and another sample (or samples) is accumulated if T_n is between a and b.

It must be emphasized that, once the quantiles are known (estimated), Eq. (1) is independent of the underlying noise distribution under both hypotheses. By assuming a Lehmann and shift of the mean alternative, P_{1k} was defined in Ref. 1 respectively, as

$$P_{1k} = F(ak)^{1+\delta_1} - F(ak-1)^{1+\delta_1} \quad \text{(Lehmann)}$$

$$P_{1k} = F(a_k - \Delta_1) - F(a_{k-1} - \Delta_1) \quad \text{(shift of the mean)}$$

where δ_1 and Δ_1 represent the selected parameter based on a priori information (say, minimum SNR of interest). It will be convenient in subsequent developments to use θ, θ_1 as parameters when discussing general properties of the SPD pertaining to both alternatives.

B. Structure of Dependent Sequential Partition Detectors

Assume $x_1, x_2, \cdots, x_n$ is a sample from a stationary q-dependent random process

with c.d.f. $F(x_1, x_2, \cdots, x_q)$. Let $\vec{a} = (a_o = -\infty, a_1, \cdots, a_{m-1}, a_m = +\infty)$ be an ordered vector of quantiles, estimated with independent samples and c.d.f. $F(x)$, subdividing the real axis into m cells. (See References 9 and 10.)

Define the random variable

$$T_{jN_o} = \sum_{i}^{N_o} \sum_{k}^{m} b_k n_{ik} \quad j = 1, 2, \cdots, n \quad , \tag{2}$$

where $N_o \geq q$ represents the number of subsamples $(x_j, x_{j+1}, \cdots, x_{j+N_o})$ summed in a decision interval.

Using the data sequence $(x_1, x_2, \cdots, x_n)$, the dependent sequential partition detector (DSPD) is constructed to test two hypotheses H_o, H_1, with stopping boundaries ($\ell n B$, $\ell n A$). Then, for independent T_{jN_o} $(j = 1, 2, \cdots, n)$,

$$T_n = \sum_{j}^{n} T_{jN_o} \quad , \tag{3}$$

and sampling continues if $\ell n B < T_n < \ell n A$ and H_o, H_1 is chosen if $T_n \leq \ell n B$, $T_n \geq \ell n A$, respectively.

The basic structure of the DSPD is shown in Fig. 1, where the quantiles $\vec{a}$ are estimated under H_o with independent samples. Once the quantiles are known, the scores b_k, $k = 1, 2, \cdots, m$ can be found if the functional form of the alternative is specified.

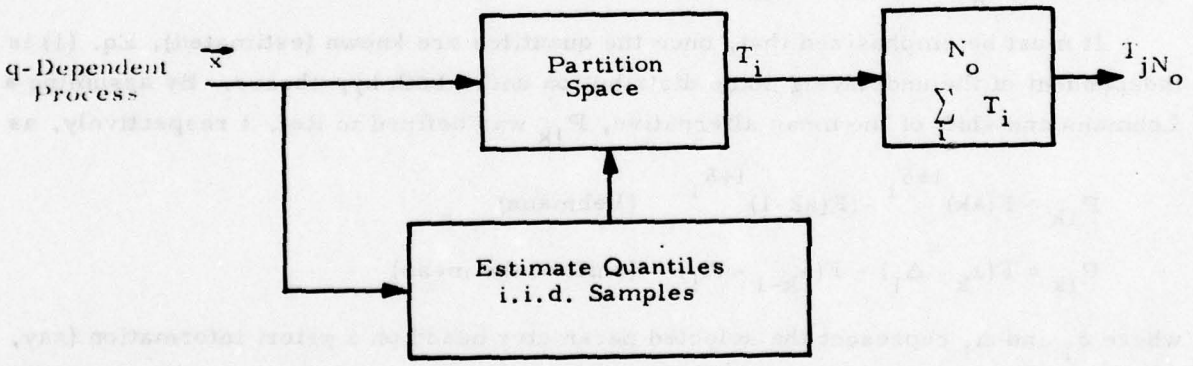


Fig. 1. Structure of DSPD.

Let the N_o-dimensional moment generating function of T_{jN_o} be defined as

$$\phi_{N_o}(\vec{t}) = E\left[\prod_{e^i}^{N_o} t_i^{T_i}\right] \quad , \tag{4}$$

where

$$T_i = \sum_k^m b_k n_{ik} \quad .$$

It is convenient to use the cumulant expansion for $\phi_{N_o}(\vec{t})$ (see Reference 2).

$$\ell n(\phi_{N_o}(\vec{t})) = \sum_i^{N_o} t_i E(T_i) + 1/2 \sum_{i,L} \lambda_{i,L} t_i t_L + O(\mu) \quad ,$$

where

$$\sum_{i,L} \lambda_{iL} t_i t_L = E(\sum_i t_i (T_i - E(T_i)))^2$$

and $O(\mu)$ represents terms greater than second order.

Since T_{jN_o} represents a sum of T_i,

$$\phi_{N_o}(\vec{t}) = \phi_{N_o}(t_1 = t,\ t_2 = t, \cdots, t_{N_o} = t) = \phi(t) \quad .$$

The cumulant expansion reduces to

$$\underset{n \to \infty}{\ell n[\phi(t)]} = t E(T_{jN_o}) + 1/2 t^2 \sigma^2 T_{jN_o}$$

Given the moment generating function, $\phi(t) = E(e^{t T_{jN_o}})$, a value of $t_D = t_o \neq 0$ can be found so that $\phi(t_o) = 1$, where

$$\underset{\substack{n \to \infty \\ \theta_1 \to 0}}{t_D(\theta) = t_o = -2 \left\{ \frac{E(T_{jN_o})}{\sigma^2 T_{jN_o}} \right\}} \quad .$$

For the class of alternatives considered, $O(\mu) \to 0$ as the selected SNR $\theta_1 \to 0$ and $n \to \infty$.

The mean of T_{jN_o} is given by

$$E(T_{jN_o}) = E(\sum_i^{N_o} \sum_k^m b_k n_{ik}) = N_o \sum_k^m b_k P_k = N_o E(T_i) \quad , \tag{5}$$

and the variance of T_{jN_o} by,

$$\sigma^2_{T_{jN_o}} = E[(\sum_{i}^{N_o}\sum_{k}^{m} b_k\, n_{ik})^2] - E^2[\sum_{i}^{N_o}\sum_{k}^{m} b_k\, n_{ik}]$$

$$= N_o\, E(T_i^2) + \sum_{i}^{N_o}\sum_{j}^{N_o}\sum_{k}^{m}\sum_{L}^{m} b_k\, b_L\, E_{|i-j|}(n_k\, n_j) - N_o^2 E^2(T_i), \qquad (6)$$

$$i \neq j$$

Where it was assumed that T_{jN_o} satisfies the strong mixing condition (Rosenblatt[3]), the joint expecatation, $E(n_{ik}\, n_{jL}) = E(n_{jL}\, n_{ik})$, is symmetric for all i, j, k and L, and the quantiles remain fixed during a decision period. Also, in Eq. (6) the dependence on the sampling rate was expressed as subscripts $|i-j|$.

The limiting value of Eq. (6) is found by letting $m \to \infty$. The second term can be asymptotically expressed as

$$\lim_{m\to\infty} \sum_{k}^{m}\sum_{L}^{m} b_k\, b_L\, E_{|i-j|}(n_k\, n_j)$$

$$= \lim_{m\to\infty} \sum_{k}^{m}\sum_{L}^{m} \ell n \left[\frac{G(a_k) - G(a_{k-1})}{(a_k - a_{k-1})} \Big/ \frac{F(a_k) - F(a_{k-1})}{(a_k - a_{k-1})} \right]$$

$$\ell n \left[\frac{G(a_L) - G(a_{L-1})}{(a_L - a_{L-1})} \Big/ \frac{F(a_L) - F(a_{L-1})}{(a_L - a_{L-1})} \right] \cdot \Big[G(a_k\, a_L) +$$

$$G(a_{k-1}, a_{L-1}) - G(a_k, a_{L-1}) - G(a_{k-1}, a_L) \Big] \longrightarrow$$

$$\int_{-\infty}^{\infty}\int_{-\infty} \Lambda(x)\ \Lambda(y)\ g_{|i-j|}(x, y)\, dxdy = E_{|i-j|}(\Lambda(x), \Lambda(y)),$$

where $\Lambda(.) = \ell n\,(g(.)/f(.))$ represents the asymptotic expression for the scores b_k, b_L, $G(a_k, a_L)$ is the joint c.d.f. under the alternative.

The limiting forms for the mean and variance of T_i are given by

$$\lim_{m\to\infty} E(T_i) = E[\Lambda(x)]$$

and

$$\lim_{m\to\infty} \sigma^2_{T_i} = (E(T_i^2) - E^2(T_i)) \to \sigma^2_{\Lambda(x)}$$

By substituting the asymptotic forms into Eq. (6) the asymptotic variance of T_{jN_o} reduces to

$$\sigma^2_{T_{jN_o}} = N_o \sigma_\Lambda^2 \left[1 + \frac{2}{N_o} \sum_{q=1}^{N_o-1} (N_o - q)\, \rho_\Lambda (q) \right],$$

where

$$\sum_{i}^{N_o} \sum_{j}^{N_o} \rho_\Lambda (i-j) = 2 \sum_{q=1}^{N_o-1} (N_o - q)\, \rho_\Lambda(q),$$

and

$$\rho_\Lambda (|\, i - j\,|) = \frac{E_{|i-j|}(\Lambda(x)\, \Lambda(y)) - E^2(\Lambda(X))}{\sigma_\Lambda^2}$$

is defined as the asymptotic correlation coefficient.

For finite m, the partition correlation coefficient is defined as

$$R_T(q) = \frac{\sum_{k}^{m} \sum_{L}^{m} b_k\, b_L\, E_q(n_k\, n_L) - E^2(T_i)}{\sigma_{T_i}^2}, \tag{7}$$

where

$$E_q(n_k\, n_L) = (G_q(a_k,\, a_L) + G_q\,(a_{k-1}\,,\, a_{L-1})$$

$$- G_q(a_k,\, a_{L-1}) - G_q(a_{k-1},\, a_L))\,.$$

Then,

$$\sigma^2_{T_{jN_o}} = \sigma^2_{T_i} N_o \left(1 + \frac{2}{N_o} \sum_{q}^{N_o-1} (N_o - q)\, R_T(q) \right).$$

For a Gaussian distribution with input correlation coefficient ψ (q), Eq. (7) can be simplified further by expanding $G(a_k,\, a_L)$ and integrating each term over ψ (q) (see Ref.2, pp. 290). The result is given by

$$E_q(n_k\, n_L) = P_k\, P_L + \int_0^{\psi} [\, g(a_k,\, a_L) + g(a_{k-1},\, a_{L-1}) -$$

$$g(a_k,\, a_{L-1}) - g(a_{k-1},\, a_L)]\, d\, \psi\, (q)\,. \tag{8}$$

where

$$g(a_k, a_L) = \frac{1}{2\pi\sigma^2} \frac{e^{-1/2 \frac{1}{\sigma^2[1-\psi 2]} (a_k^2 + a_L^2 - 2a_k a_L \psi)}}{\sqrt{1-\psi^2}}$$

and $P_k = G(a_k) - G(a_{k-1}) = F(a_k - \Delta) - F(a_{k-1} - \Delta)$, under a shift of the mean alternative.

Example 1

Let $m = 2$, $\Delta = 0$, $b_1 = -b_2$, $\sigma^2 = 1$, and the optimum quantiles are given by[1]

$$\begin{cases} a_o = -\infty \\ a_1 = 0 \\ a_2 = +\infty \end{cases} .$$

Then, $P_k = P_L = 1/2$

$$E_q(n_1 n_1) = E_q(n_2 n_2) = 1/4 + \int_0^{\psi} \frac{1}{2\pi \sqrt{1-\psi^2}} d\psi$$

$$= 1/4 + \frac{1}{2\pi} \arcsin \psi(q),$$

$$E_q(n_1 n_2) = E_q(n_2 n_1) = 1/4 - \frac{1}{2\pi} \arcsin \psi(q)$$

and the partition correlation coefficient reduces to

$$R_T(q) = \frac{2}{\pi} \arcsin \psi(q).$$

Equation (8) was evaluated using numerical analysis techniques for $m = 2, 4$ and 6; $\Delta = 0$, 1 and $0 \leq \psi(q) < 1$.

The results are given in Figures 2 and 3. Note that $\psi(q) \geq R_T(q)$ for all m, and as $m \to \infty$ $R_T(q) \to \psi(q)$ very rapidly. Therefore, from these results for a Gaussian process, it can be concluded that $\rho_\Lambda(q)$ gives an upper bound for $\sigma^2_{T|N_o}$. Unfortunately, for other distributions the evaluation of Eq. (7) is difficult, since a two-dimensional integration is required. However, using a series expansion[4] and assuming small $\psi(q)$ for the Rayleigh process under a Lehmann alternative, it was shown that $\rho_\Lambda(q) \geq R_T(q)$ for $m = 2$ and 4.

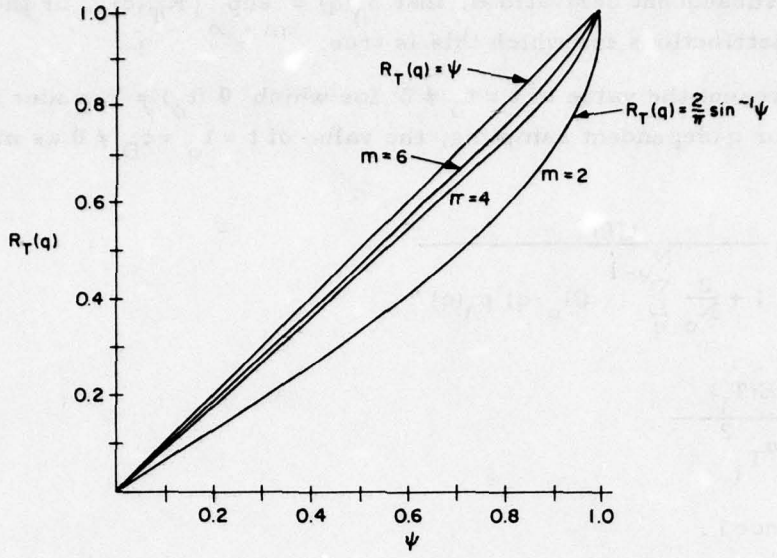


Fig. 2. Evaluation of $R_T(q)$ for m = 2, 4, 6 and $\Delta = 0$.

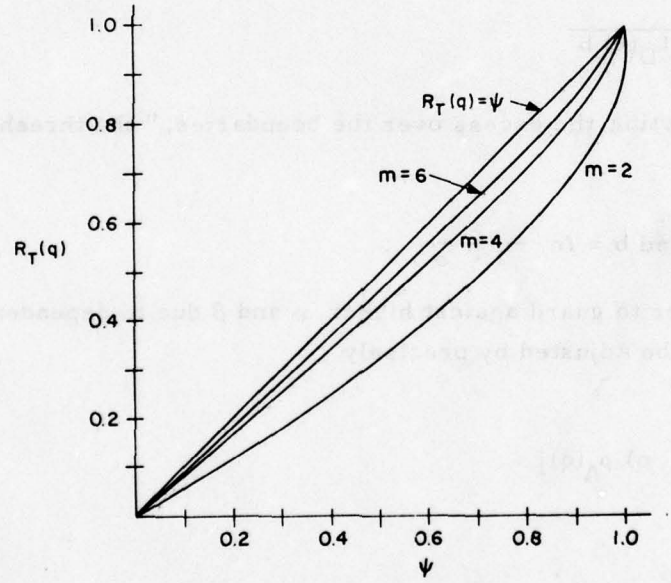


Fig. 3. Evaluation of $R_T(q)$ for m = 2, 4, 6 and $\Delta = 1$.

From the above results, and since $\lim_{m \to \infty} R_T(q) \to \rho_\Lambda(q)$, it will be assumed (without proof) in subsequent derivations, that $\rho_\Lambda(q) = \sup_{m \leq \infty} [R_T(q)]$, or the subsequent results hold for distributions for which this is true.

Let $t(\theta)$ represent the value of $t = t_o \neq 0$ for which $\emptyset(t_o) = 1$, under independent sampling, then, for q-dependent sampling, the value of $t = t_o = t_D \neq 0$ as $m \to \infty$ is given by

$$t_o = t_D(\theta) \underset{m \to \infty}{=} \frac{t(\theta)}{1 + \frac{2}{N_o} \sum_{q}^{N_o - 1} (N_o - q)\, \rho_\Lambda(q)} , \tag{9}$$

where

$$t(\theta) \underset{\theta_1 \to 0}{=} -2 \frac{E(T_i)}{\sigma_{T_i}^2}$$

is given in Reference 1.

Notice the effects of dependence (rapid sampling) only show up in the solution of $\emptyset(t) = 1$. From the fundamental identity[5] the probability of detection becomes

$$P(\theta) = \frac{1 - e^{t_D(\theta)\, b}}{e^{t_D(\theta)\, a} - e^{t_D(\theta)\, b}} \tag{10}$$

Also, from Wald, "neglecting the excess over the boundaries," the thresholds are given by,

$$a = \ell n \frac{1 - \beta}{\alpha} \quad \text{and} \quad b = \ell n \frac{\beta}{1 - \alpha} .$$

Therefore, in order to guard against higher α and β due to dependence, the thresholds a and b must be adjusted by precisely

$$\left[1 + \frac{2}{N_o} \sum_{q}^{N_o - 1} (N_o - q)\, \rho_\Lambda(q)\right] .$$

Define

$$a' = \left[1 + \frac{2}{N_o} \sum_{q}^{N_o - 1} (N_o - q)\, \rho_\Lambda(q)\right] a$$

and

$$b' = \left[1 + \frac{2}{N_o} \sum_{q}^{N_o - 1} (N_o - q)\, \rho_\Lambda(q)\right] b$$

as the new thresholds needed for a q-dependent process to assure the same protection against α and β as in the independent case.

The average sample number[6] for a q-dependent process (ASND) with adjusted thresholds a', b' under a Lehmann or shift of the mean alternative becomes then,

$$\text{ASND} = \frac{\left[1 + \frac{2}{N_o}\sum_{q}^{N_0-1}(N_o-q)\,\rho_{\Lambda}(q)\right]}{N_o\,E(T_i)} \frac{b(e^{t(\theta)a}-1) + a(1-e^{t(\theta)b})}{e^{t(\theta)a} - e^{t(\theta)b}}, \quad E(T_i) \neq 0 \tag{11}$$

and

$$\text{ASND} = \frac{-ab\left[1 + \frac{2}{N_o}\sum_{q}^{N_o-1}(N_o - q)\,\rho_{\Lambda}(q)\right]}{N_o\sigma^2_{T_i}}, \quad E(T_i) = 0. \tag{12}$$

Define t_I/t_D as the ratio of time required for independent observations to the total time needed for a decision interval, including enough delay to assure independent samples $[T_{jN_o}, T_{j+1}, N_o]$.

It follows that

$$\frac{t_I}{t_D} = \frac{r_s}{N_o + r_s - 1}, \tag{13}$$

where r_s/Δ' is the sampling rate, Δ' is the time between independent observations and r_s is an integer. For $r_s = 1$, $1/\Delta'$ is the sampling rate for independent observations. If the sampling rate is increased r_s times, $r_s - 1$ samples are skipped after each decision interval, T_{jN_o}.

Table I gives typical threshold adjustment factors for a Gaussian process based on $\rho_{\Lambda}(q)$ as a function of r_s and N_o.

TABLE I. Threshold adjustment factors for a Gaussian process.

N_o	r_s (sampling rate)		
	2	3	4
2	1.2	-	1.5
3	-	1.6	-
4	1.4	-	2.0
6	-	1.9	-
8	1.5	-	2.4
∞	1.6	2.2	2.8

Note

Since $\rho_\Lambda(q) \geq R_T(q)$, the adjusted thresholds a', b', give an upper bound on the errors α and β.

It should also be pointed out that in order to calculate the new thresholds a', b', in practical implementation, the form and value of $\rho_\Lambda(q)$ must be known. This limitation could be removed by estimating $\rho_\Lambda(q)$ at, say, a few sampling rates and setting up thresholds a', b' which will bound the errors,

$$\alpha' \leq \alpha$$

$$\beta' \leq \beta$$

as long as the predicted dependence is not exceeded.

There does not seem to be any general technique which will eliminate the need to know $\rho_\Lambda(q)$ in all cases.

C. Efficiency Calculations for Dependent Sequential Partition Detectors

Let N_o (block size) represent the number of samples summed in a decision interval. If the sampling rate is such that $r_s = N_o$, there can be no increase in ASND solely from accumulating the samples in blocks. The reason for this is that all dependent samples fall into the independent block size N_I, neglecting for the moment the additional delay needed between dependent blocks. However, if $N_o > r_s N_I$, then grouping the samples in blocks can only increase ASND, except in a special case. Suppose n_I is the number of groups needed to terminate a test when the samples are independent and blocks are of size N_I. If n_I is an integral multiple of N_o, then grouping in blocks of N_o when samples are dependent will not in itself increase ASND. From the discussion in the last section, dependence effects could be removed from $P(\theta)$ by increasing the thresholds to a' and b'. In the subsequent discussion, assume this has been done.

In general, the amount of protection afforded by sampling in blocks depends upon N_o, m, and θ. Since θ_1 can be made as small as desired, and usually N_o is bounded, $P(\theta)$ will not substantially change (θ_1 small) when sampling in blocks. Therefore, the probability of detection $P(\theta)$ when making efficiency calculations will be assumed constant. Also, for sufficiently small $\theta_1 (n \to \infty)$, the effect of grouping on the ASND is negligible.[6]

In order to compare the relative efficiencies of two test using dependent sampling, the ASND will be based on the time it takes test T_1 to terminate compared with test T_2.

Let t_I and t_D represent the time between data samples for SPD and DSPD, respectively. Then, the efficiency of DSPD with respect to SPD is defined by

$$\text{Efficiency} = \mathcal{E}(T_1/T) = \frac{\text{ASN} \cdot t_I}{\text{ASND} \cdot t_D} = \frac{\text{ASN}}{\text{ASND}} \quad \frac{r_s}{N_o + r_s - 1} \quad ,$$

where Eq. (13) was employed using an independent sampling rate of $1/\Delta'$.

1. Lehmann Alternative Case

Let $N_o = r_s$ and define, for fixed α β ,

$$\mathcal{E}\left[\frac{T_{L\infty}}{T_{L\,Dm}}\right] = \frac{\text{ASN}L\infty}{\text{ASNDL}} \quad \frac{r_s}{N_o + r_s - 1}$$

$$= \frac{\sum_k^m \dfrac{[F(a_k)\ \ell n\, F(a_k) - F(a_{k-1})\ \ell n\, F(a_{k-1})]^2}{F(a_k) - F(a_{k-1})}}{[1 + 2/N_o \sum_q^{N_o - 1} (N_o - q)\ \rho_\Lambda(q)]} \quad \frac{N_o r_s}{N_o + r_s - 1} \quad ,$$

where $T_{L\infty}$ is defined as the optimum test, $m \to \infty$, and [1]

$$\lim_{m \to \infty} \sum_k^m \frac{[F(a_k)\ \ell n\, F(a_k) - F(a_{k-1})\ \ell n\, F(a_{k-1})]^2}{F(a_k) - F(a_{k-1})} = 1,$$

Then,

$$\mathcal{E}_{m \to \infty}\left[\frac{T_{L\infty}}{T_{LDm}}\right] = \frac{1}{[1 + \frac{2}{N_o} \sum_q^{N_o - 1} (N_o - q)\rho_\Lambda(q)]} \quad \frac{N_o r_s}{N_o + r_s - 1}$$

Now, let $r_s \to \infty$, $N_o \to \infty$.

$$\mathcal{E}_{\substack{m \to \infty \\ r_s \to \infty \\ N_o \to \infty}}\left[\frac{T_{L\infty}}{T_{LDm}}\right] \to \frac{\Delta'/2}{\frac{\Delta'}{r_s} \sum_q^{N_o - 1} \rho_\Lambda(q)} \to \frac{\Delta'/2}{\int_0^{\Delta'} \rho_\Lambda(\mu)\, d\mu} \tag{14}$$

Kassam and Thomas[7] obtained the same result for a fixed sample sign test in Gaussian noise. However, the above result also applies to dependent sequential

partition detectors and is independent of the distribution under the Lehmann alternative, and for $m > 2$ represents a generalization of the sign ($m = 2$) test.

If $N_o > r_s$ and r_s is fixed, the above results apply as long as $N_o < \infty$, and if it can be assumed that as $n \to \infty$, $\theta_1 \to 0$, effect of grouping is negligible.

Example 2

Let the noise distribution be a Gauss-Markov process.[8] Then,

$$\rho_\Lambda(\mu) = e^{-k/\Delta' |\mu|}$$

and for independent samples $\mu = \Delta'$, and if $k = 3$, the correlation between two adjacent samples is .05. Then, from Eq. (14)

$$\underset{\substack{m\to\infty \\ r_s\to\infty \\ N_o\to\infty}}{\mathcal{E}}\left[\frac{T_{L\infty}}{T_{LDm}}\right] \to 1.5 ,$$

which is a significant improvement in efficiency for the DSPD.

Figure 4 compares efficiencies for $m = 2$ and $m = 8$ when $r_s = 2$ and N_o is increased. The curve for $m = 8$ depicts the improvement in efficiency over the sign ($m = 2$) test under this condition. The improvement in efficiency for $m = 8$ over the sign ($m = 2$) test as $N_o \to \infty$, $r_s \to \infty$, approaches

$$\underset{\substack{r_s\to\infty \\ N_o\to\infty}}{\mathcal{E}}\left[\frac{T_{L2}}{T_{LD8}}\right] \to 2.2$$

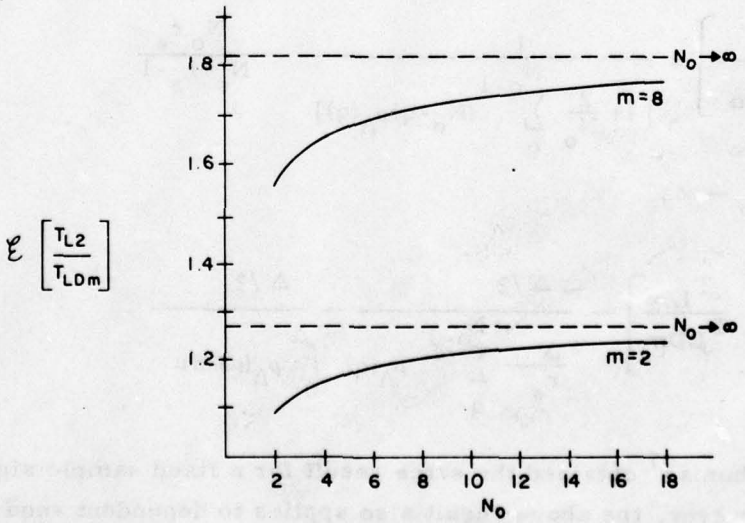


Fig. 4. Efficiency of DSPD vs SPD($m = 2$).

2. Shift of the Mean Alternative Case

Let

$$\mathcal{E}\left[\frac{T_{s\infty}}{T_{sDm}}\right] = \frac{ASNS\infty}{ASNDS} \quad \frac{r_s}{N_o + r_s - 1}$$

be the efficiency for the DSPD compared to the optimum ($m \to \infty$) test under a shift of the mean alternative when α and β are fixed.
Then,

$$\mathcal{E}\left[\frac{T_{s\infty}}{T_{sDm}}\right] = \frac{\sum_{k}^{m} \frac{[f(a_{k-1}) - f(a_k)]^2}{F(a_k) - F(a_{k-1})}}{\left[1 + \frac{2}{N_o}\sum_{q}^{N_o-1}(N_o - q)\rho_{\Lambda}(q)\right]} \quad \frac{\frac{N_o r_s}{N_o + r_s - 1}}{\int_{-\infty}^{\infty}\{\frac{f'(x)}{f(x)}\}^2 f(x)\,dx}$$

and as $m \to \infty$

$$\lim_{m\to\infty}\mathcal{E}\left[\frac{T_{s\infty}}{T_{sDm}}\right] = \frac{1}{\left[1 + \frac{2}{N_o}\sum_{q}^{N_o-1}(N_o - q)\rho_{\Lambda}(q)\right]} \quad \frac{N_o r_s}{N_o + r_s - 1} .$$

Also assuming a Gauss-Markov process as in Example 2, the efficiency approaches

$$\lim_{\substack{m\to\infty \\ r_s\to\infty \\ N_o\to\infty}} \mathcal{E}\left[\frac{T_{s\infty}}{T_{sDm}}\right] \to 1.5$$

as in the Lehmann case.

Note that F(x) must be specified in the shift case except when $m \to \infty$. In general, the efficiencies for two non-optimum ($m < \infty$) tests can be compared,

$$\mathcal{E}\left[\frac{T_{sm_1}}{T_{sDm_2}}\right] = \frac{\sum_{k}^{m_2} \frac{[f(a_{k-1}) - f(a_k)]^2}{F(a_k) - F(a_{k-1})} \quad \frac{N_o r_s}{N_o r_s - 1}}{\left[1 + \frac{2}{N_o}\sum_{q}^{N_o-1}(N_o - q)\,p_{\Lambda}(q)\right]\sum_{k}^{m_1}\frac{[f(a_{k-1}) - f(a_k)]^2}{F(a_k) - F(a_{k-1})}}$$

D. Conclusions

The effects of dependent sampling on sequential partition detectors (SPD) were considered in this report. First, the theory of SPD was extended to include q-dimensional processes by introducing a linear transformation from a q-dimensional space to a one dimensional space. Then using Wald's fundamental identity the probability of detection and average sample number for the DSPD were derived. It was shown how to adjust the thresholds (a, b) in order to obtain the same error probabilities as in the independent sampling case. Next, for the adjusted thresholds (a', b') efficiency calculations were made under Lehmann and shift of the mean alternatives. The interesting results were that both alternatives gave improved efficiencies under dependent sampling.

Joint Services Technical Advisory Committee
F44620-74-C-0056

R. F. Dwyer and L. Kurz

Naval Underwater Systems Center

REFERENCES

1. R. Dwyer and L. Kurz, "Sequential Partition Detectors," Cybernetics (to appear).
2. J. Cramer, Mathematical Methods of Statistics, Princeton University Press, (1951).
3. M. Rosenblatt, "Independence and Dependence," Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 11 (1961).
4. J. F. Barrett and D. G. Lampard, "An Expansion for Some Second-Order Probability Distributions and Its Application to Noise Problems," IRE Trans. Inform. Theory, Vol. IT-1, pp. 10-15 (March, 1955).
5. A. Wald, Sequential Analysis, (New York: John Wiley and Sons 1947).
6. R. F. Dwyer, "Sequential Partition Detectors with Applications," Ph. D. Dissertation, Polytechnic Institute of New York (June 1976).
7. S. Kassam and J. Thomas, "A Class of Nonparametric Detectors for Dependent Input Data," IEEE Trans. on Inform. Theory, Vol. IT-21, No. 4 (July 1975).
8. M. Woinsky and L. Kurz, "Nonparametric Detection Using Dependent Samples," IEEE Trans. on Inform. Theory (Corresp.), Vol. IT-16, pp. 355-358 (May, 1970).
9. P. Kersten and L. Kurz, "Robustized Vector Robbins-Monro Algorithm with Applications to M-Interval Detection," Information Sciences, (October 1976).
10. P. Kersten and L. Kurz, "Bivariate M-Interval Classifiers with Application to Edge Detection," Information and Control (June 1977).

SEQUENTIAL PARTITION DETECTORS IN LARGE SIGNAL AND IMPULSIVE NOISE ENVIRONMENTS

R.F. Dwyer and L. Kurz

In the previous report, the concept of partition detection was extended to include sequential detection[1] and certain types of dependence.[2] In this report, the theory of sequential partition detectors is extended to include large signals and impulsive noise. The notation and definitions of Refs. 1 and 2 are used in this report.

The cases of large signals in noise and small signals in impulsive noise will be treated separately. In most practical situations, these are the cases of interest. However, there can be some degree of overlap, i.e., moderate signal strengths in impulsive noise, without having to redesign the sequential detector.

The performance, on the other hand, is difficult to predict since it depends upon the severity of the impulsive noise, its frequency and time of occurrence. For example, if an impulse occurred, when the statistic, T_n, was near one of its boundaries (a,b), with sufficient strength to terminate the test, large errors can be expected. However, if T_n is between its boundaries when an impulse occurs and the test does not terminate immediately, it is not unreasonable to assume that for a sufficient number of samples, without an impulse occurring, the test could respond and make the correct decision within its designed error probabilities.

From this example, it is quite evident that sequential tests are more sensitive (i.e., can make the wrong decision) to impulsive noise than fixed sample tests, since, usually, a fixed sample test will have time to "average out," the impulsive interference, with everything else considered equal.

Also notice, that the partition structure itself is somewhat insensitive to impulsive noise. That is, for a particular partition, say m = 2, all interfering noise levels are given the same value, not unlike a hard clipper, and, depending upon the design, each impulse contributes only a small amount to the total.[3]

In the next section, sequential partition detectors are designed to operate in large signal environments where the noise is undefined except that the quantiles will be assumed known. The basic equations for sequential analysis are still valid for this case, however, the approximations made in finding a solution to the moment generating function in the preceding work (Refs. 1 and 2) are not. Also, neglecting the excess over the boundaries would not be appropriate in this case, since both $|E(t_i)|$ and $\sigma^2_{T_i}$ at (θ_0, θ_1) are expected to be large and, therefore, the ASN small.

In the impulsive noise case, a reduced partition space will be introduced which discards (i.e., assigns weight zero), to samples falling outside the space. This is

effectively a partition correlator which only accepts samples falling within its defined range. Assuming that impulsive noise is composed of high amplitude, large variance impulses, occurring randomly and the signals of interest are of low energy in noise of small variance occurring during a decision interval, the space can then be partitioned into two disjoint regions. A signal space and an impulsive noise space will be defined in R_1. An impulse falling into the signal space or a signal sample falling into the impulsive space cannot be discerned further. The main concern however, will be to reduce the sensitivity to impulsive noise using the same simple structure of the SPD under the Lehmann and shift alternatives. The price paid for this insensitivity will be generally, an increased ASN. A similar approach was introduced by Rappaport and Kurz for digital signalling in nongaussian noise problems.[4]

A. A Sequential Partition Detector for Large SNR Cases

For some situations, the small SNR model of the SPD developed in the previous work may not be applicable. In active sonar, for example, detection of a target is desired in only a few "pings" (samples) when the energy from the echo return is large. Therefore, under these circumstances, another approach is taken.

Assume $x_1, x_2, \cdots, x_n$ is a random sample of i.i.d.r.v. with c.d.f. F(x) and a is a vector of constants subdividing the real axis into m cells as discussed previously. Also, let

$$T_n = \sum_i^n T_i = \sum_i^n \sum_k^m b_k n_{ik}$$

be the nonparametric test statistic and n_{ik}, b_k are as defined previously.

Recall that the fundamental identity is given by,

$$E(e^{T_n t} \phi(t)^{-n}) = 1 \quad , \tag{1}$$

where

$$\phi(t) = \sum_k^m P_k e^{b_k t} = 1 \tag{2}$$

The following is a modified version of Wald's[5] derivation. Assume the scores b_k only take on integral multiple values of some constant d. Since the scores are functions of the likelihood ratio, this assumption restricts somewhat the quantile locations and also the selected alternative parameter θ_1. Therefore, for the Lehmann and shift alternatives, the efficiency will be reduced from their optimum values. Fortunately, this is not serious, since some loss in efficiency can be tolerated in the high SNR case.

However, for a shift alternative and m = 2 under Gaussian noise, no restrictions are imposed by assuming the scores are multiple values of d. In general, finding the common factor d for m> 2 is difficult and is, therefore, not a very useful technique. However, for those cases where d can be found, it does provide a means of obtaining $P(\theta)$ and $E(n/\theta)$ without approximation. Later, an example will be given, for m = 2 under a shift alternative, comparing the results using the approximation for $E(n/\theta)$ and the exact technique for θ large.

Again, the roots of Eq. (2) must be found. Let $u = e^t$, then

$$\sum_k^m P_k u^{b_k} = 1 \tag{3}$$

There are m' roots to Eq. (3), $u_i(\theta)$ $(i = 1, 2, \cdots, m')$, and assuming that they are all unique, Eq. (1) reduces to

$$E(u_i(\theta)^{T_n}) = 1, \quad (i = 1, 2, \cdots, m')$$

At termination, T_n can take on only a finite set of values $(c_1, c_2, \cdots, c_m)$ which are determined from the modified thresholds (da, db) and the values of b_k. If, for example, m = 2, da = 5, db = -5, and $b_1 = -1$, $b_2 = +1$, then the test can only terminate at ± 5.

Therefore,

$$\sum_k^{m'} P(T_n = c_k/\theta)\, u_i(\theta)^{c_k} = 1, \quad (i = 1, 2, \cdots, m')$$

or in the determinant form

$$P(T_n = c_k/\theta) = \frac{\begin{vmatrix} u_1(\theta)^{c_1} & \cdots & u_1(\theta)^{c_{k-1}} & 1 & u_1(\theta)^{c_{k-1}} & \cdots & u_1(\theta)^{c_{m'}} \\ \cdot & & \cdot & \cdot & \cdot & & \cdot \\ u_{m'}(\theta)^{c_1} & \cdots & u_{m'}(\theta)^{c_{k-1}} & 1 & u_{m'}(\theta)^{c_{k+1}} & \cdots & u_{m'}(\theta)^{c_{m'}} \end{vmatrix}}{\begin{vmatrix} u_1(\theta)^{c_1} & \cdots & \cdots & u_1(\theta)^{c_{m'}} \\ \cdot & & & \cdot \\ u_{m'}^{c_1} & \cdots & \cdots & u_{m'}(\theta)^{c_{m'}} \end{vmatrix}}$$

Then the probability of detection is the sum over all k for which $c_k \geq ad$,

$$P(\theta) = \sum_k P(T_n = c_k/\theta) \quad , \tag{4}$$

and

$$E(n/\theta) = \frac{d}{E(T_i)} \sum_{k}^{m'} c_k P(T_n = c_k/\theta) \tag{5}$$

where both Eq. (4) and Eq. (5) are exact in that the excess over the boundary is not neglected.

These equations are general, but unfortunately, very cumbersome for comparison purposes. Some simplification in the evaluation of these equations was given by Girshich[6] but they are still formidable for practical interpretation. A practical technique for evaluating the ASN and the probability distribution P(n) was introduced by Proakis.[7] The SPD is represented, in this technique, as a Markov process with absorbing boundaries a,b. The states of the Markov process will correspond to possible values of T_n.

First, an example will be given, using Wald's procedure, for the case when m = 2 under a shift alternative. Next, P(n) and ASN for SPD will be calculated for m = 2, 4, 6 using Proakis' technique under a shift alternative. Then the results of a simulation in Gaussian noise, N(0, 1) under a shift alternative, will be compared for m = 2, 4, and 6, to Proakis' technique.

Example 1: Bernoulli Trials

Let m = 2 and assume $b_1 = -1$ and $b_2 = +1$. Then,

$$\sum_{k}^{2} P_k e^{tb_k} = P_i e^{-t} + P_2 e^{t} = 1 \tag{6}$$

Since $P_2 = 1 - P_1$, the roots of Eq. (6) are

$$u_1 = \frac{P_1}{1 - P_1}, \quad u_2 = 1 \quad .$$

At this point, the value for t for which $\phi(t) = 1$ can be found,

$$t = -\ell n\left(\frac{P_1}{1 - P_1}\right) \quad .$$

Then, using Eq. (4) and Eq. (5) and noting that $P_1 = \delta_1(\theta)$ the probability of detection and the ASN can be found from,

$$P(\theta) = \frac{\begin{vmatrix} \left\{\frac{P_1(\theta)}{1-P_1(\theta)}\right\}^{db} & 1 \\ 1 & 1 \end{vmatrix}}{\begin{vmatrix} \left\{\frac{P_1(\theta)}{1-P_1(\theta)}\right\}^{db} & \left\{\frac{P_1(\theta)}{1-P_1(\theta)}\right\}^{-da} \\ 1 & 1 \end{vmatrix}}$$

$$= \frac{1 - \left\{\frac{P_1(\theta)}{1-P_1(\theta)}\right\}^{db}}{\left\{\frac{P_1(\theta)}{1-P_1(\theta)}\right\}^{-da} - \left\{\frac{P_1(\theta)}{1-P_1(\theta)}\right\}^{db}} = \frac{1 - e^{-bdt(\theta)}}{e^{dat(\theta)} - e^{-bdt(\theta)}} \tag{7}$$

and

$$E(n/\theta) = \frac{d}{E(T_i)} \frac{-b(e^{dat(\theta)} - 1) + a(1 - e^{-bdt(\theta)})}{e^{dat(\theta)} - e^{-bdt(\theta)}} \quad , \quad E(T_i) \neq 0 \tag{8}$$

$$= \frac{-ba}{\sigma^2_{T_i}} d^2 \quad , \quad E(T_i) = 0$$

where,

$$E(T_i) = \sum_k^2 b_k P_k = 1 - 2P_1(\theta)$$

and

$$\sigma^2_{T_i} = \sum_k^2 b_k^2 P_k - (E(T_i))^2 = 1 - (1 - 2P_1(\theta))^2 \quad .$$

Note the constant d was only expressed in the thresholds.

Figure 1 compares the ASN and the power function $P(\Delta)$, using Wald's method, with the SPD estimate for a selected alternative of $\Delta_1 = 1$ and $m = 2$. The SPD estimates of ASN and $P(\Delta)$ agree very well with the exact values calculated by Wald's method over the range $0 \leq \Delta \leq \Delta_1$. Figures 2, 3 and 4 show the exact probability distribution P(n) for $\Delta = .5$, $m = 2, 4$ and 6; $\Delta = 1$, $m = 2, 4$ and $\Delta = 2$, $m = 2, 4$ and 6, respectively. The exact ASN was obtained from the figures by calculating the probability of terminating the test at each n. Table I compares the ASN obtained from Proakis' exact method and the ASN measured from a simulation of the SPD model in Gaussian noise, N(0, 1), with the SPD estimate. It is concluded from the results in the figures, that the approximation

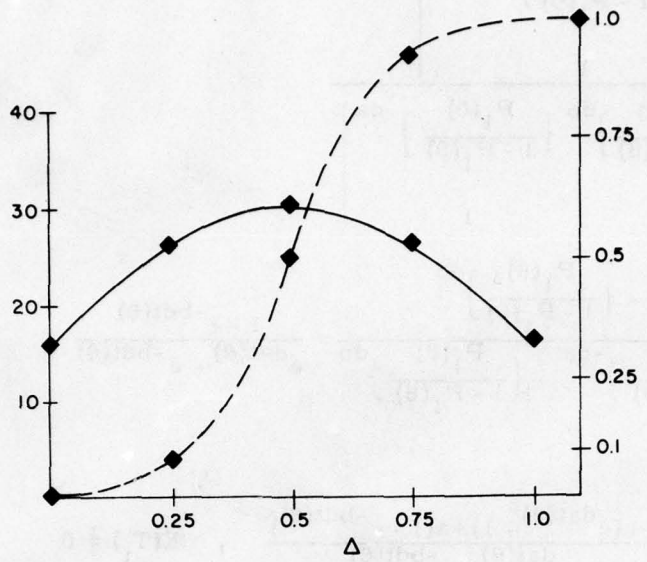


Fig. 1. Exact power and ASN functions vs. SPD estimate for m = 2.

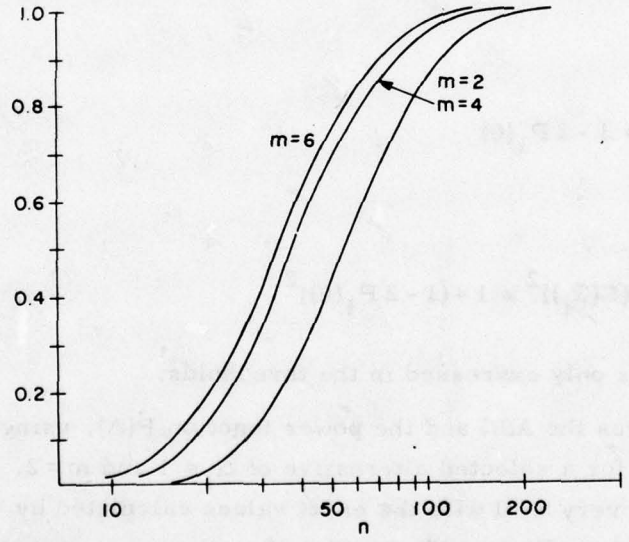


Fig. 2. Exact probability distribution P(n) for SPD, Δ = .5 Gaussian noise N(0, 1).

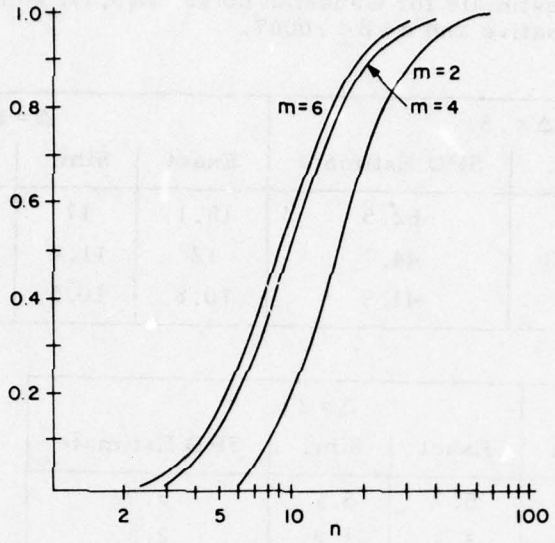


Fig. 3. Exact probability distribution P(n) for SPD, $\Delta = 1$ Gaussian noise N(0, 1).

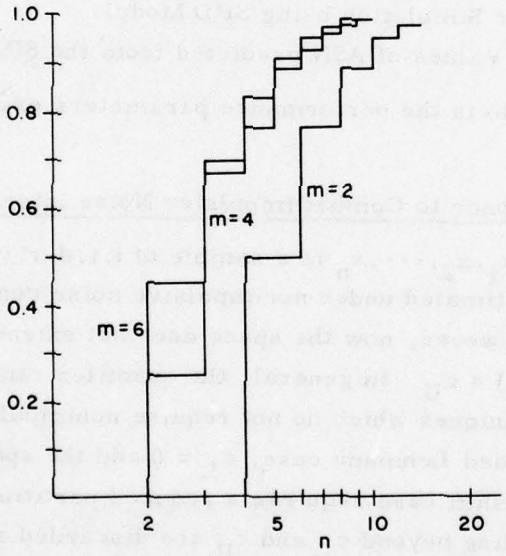


Fig. 4. Exact probability distribution P(n) for SPD, $\Delta = 2$ Gaussian noise N(0, 1).

TABLE I. ASN comparison of exact and simulation results with the SPD estimate for Gaussian noise, N(0, 1), under a shift alternative and $\alpha = \beta \leq .0067$.

	$\Delta = .5$			$\Delta = 1$		
m	Exact	Sim.	SPD Estimate	Exact	Sim.	SPD Estimate
2	64.3	59	62.5	18.1	17	15.3
4	45.7	43.3	44.7	12	11.4	11.2
6	40.9	35	41.9	10.8	10.5	10.5

	$\Delta = 2$		
m	Exact	Sim.	SPD Estimate
2	5.7	5.7	3.9
4	3.3	3.2	2.8
6	3	2.8	2.6

Exact - Proakis' Technique;

Sim. - Computer Simulation using SPD Model;

SPD Estimate - Values of ASN predicted from the SPD equations.

technique for the SPD predicts the performance parameters exceedingly well, even for moderately high SNR.

B. *A Reduced Partition Space to Combat Impulsive Noise Interference*

As before, assume $x_1, x_2, \cdots, x_n$ is a sample of i.i.d.r.v. with c.d.f. $F(x)$ and a is a vector of constants estimated under nonimpulsive noise conditions subdividing the real axis into m cells. However, now the space does not extend to $\pm\infty$ but is reduced, i.e., $F(a_o) = c_L$ and $F(a_m) = c_U$. In general, the quantiles can be estimated using robustized estimation techniques which do not require nonimpulsive noise conditions for estimation.[8] In the one-sided Lehmann case, $c_L = 0$ and the space is only restricted to the right. The two-sided shift case requires a reduced partition at both ends of the distribution. Samples falling beyond c_L and c_U are discarded and not considered in the sequential test. In this way, the SPD becomes insensitive to high amplitude impulse noise interference. Since the amplitude and frequency of occurrence of impulsive noise are assumed random, the reduced partition structure represents the best compromise in suppressing the contribution from interference and maintaining the simple structure of the SPD. However, the amplitudes of the impulsive noise are distributed throughout the space and some portion of the samples will fall within the range of the reduced

partition. Therefore, the reduced partition structure acts much like a partition correlator or matched partition filter being sensitive to samples only falling within its defined range.

Notice that now the SPD is independent from the severity of the impulsive noise. High amplitudes are discarded and low amplitudes are treated as nonimpulsive noise but are of no consequence due to the insensitivity of the partition structure.[3,8,9]

The question to be considered now is, how does the reduced partition structure affect the performance of the SPD in nonimpulsive noise environments?

Define n_k as the number of observations falling into the interval $I[a_{k-1} < x \leq a_k; k = 1, 2, \cdots, m; i = 1, 2, \cdots, n]$.

Let $P_{\nu k} = P_\nu[a_{k-\rho} < x_i \leq a_k] = F_\nu(a_k) - F_\nu(a_{k-1})$, $(\nu = 0, 1)$ represent the probability of a sample, x_i, falling into the k-th interval under the hypothesis or alternative under the constraint that

$$\sum_k^m P_{\nu k} = c_U - c_L, \quad \text{and} \quad \sum_k^m n_k = n \quad .$$

Following the procedure of Ref. 17

$$T_n = \sum_i^n T_i = \sum_i^n \sum_{k=0}^m b_k n_{ik} \quad . \tag{9}$$

Notice that now, two additional quantiles, $F(a_o) = c_L$ and $F(a_m) = c_U$ must be estimated in establishing Equation (9).

Assuming the same conditions hold as in Ref. 1, the mean and variance of T_i must be found to obtain the solution for the moment generating function.

Then,

$$E(T_i) = \sum_{k=0}^m b_k P_k = \theta_1 \sum_{k=0}^m A_k + (\theta\,\theta_1 - \tfrac{1}{2}\theta_1^2) \sum_{k=0}^m \frac{A_k^2}{P_{ok}} \tag{10}$$

and

$$\sigma_{T_i}^2 = \theta_1^2 \sum_{k=0}^m \frac{A_k^2}{P_{ok}} \tag{11}$$

where

$$\sum_{k=0}^{m} A_k = \sum_{k=0}^{m} \left(F(a_k) \ln F(a_k) - F(a_{k-1}) \ln F(a_{k-1})\right)$$

$$= F(a_o) \ln F(a_o)$$

$$+ F(a_1) \ln F(a_1) - F(a_o) \ln F(a_o)$$

$$\vdots$$

$$+ F(a_m) \ln F(a_m) - F(a_{m-1}) \ln F(a_{m-1})$$

$$= c_U \ln c_U \quad .$$

The solution for $t_o(\theta)$ takes a rather complicated form,

$$t_o(\theta) = (1 - 2\theta/\theta_1) - 2(c_U \ln c_U)\Big/\theta_1 \sum_{k=0}^{m} A_k^2/P_{ok} \tag{12}$$

However, usually $c_L \approx 0$ and $c_U \approx 1$ and Eqs. (10), (11) and (12) reduce to those given in Reference 1. For example, under a two-sided shift alternative, for $c_L = .0001$ and $c_U = .9999$, there is only a slight change in the performance parameters and these differences can be neglected for practical considerations.

In order to compare the performance under impulsive noise interference, a simulation was conducted for the shift case. The quantiles were chosen under Gaussian noise, N(0, 1), and fixed. The interference was also Gaussian, but with large variance, N(0, 10), which occurred 20 percent of the time. Table II compares the ASN and the number of errors made in 40 trials for both the normal and reduced partition space with $c_L = .0001$ and $c_U = .9999$. The space was partitioned into two disjoint regions; from -3.62 to 3.62 represents the signal space, and from $-\infty$ to -3.62 and from 3.62 to $+\infty$ represents the impulsive noise space. Two signals ($\Delta = \Delta_1 = .5$ and $\Delta = \Delta_1 = 1$) were considered in the simulation for m = 2, 4 and 6. For all cases, the reduced partition space rejected the interference and made the correct decision. It is evident that the sensitivity (measured in errors made) to impulsive interference increased with m under a normal partition structure; whereas the reduced partition structure's sensitivity remained constant at the expense of higher ASN.

It should also be pointed out that for one-sided alternatives, impulsive noise is more troublesome and the reduced partition space structure is, therefore, more effective in combating interference.

TABLE II. A simulation of a SPD in Gaussian N(0, 1), and Impulsive Gaussian N(0, 10) noise.

Normal Partition Space					Reduced Partition Space				
	Δ = .5		Δ = 1			Δ = .5		Δ = 1	
m = 2	ASN	Errors	ASN	Errors	m = 2	ASN	Errors	ASN	Errors
No Impulsive Interference	59.1	0	16.95	0	No Impulsive Interference	59.1	0	16.95	0
Impulsive Interference	67	0	20.45	0	Impulsive Interference	68.55	0	21.55	0
	Δ = .5		Δ = 1			Δ = .5		Δ = 1	
m = 4	ASN	Errors	ASN	Errors	m = 4	ASN	Errors	ASN	Errors
No Impulsive Interference	44.28	0	12.58	0	No Impulsive Interference	44.28	0	12.58	0
Impulsive Interference	47.1	1	13.5	1	Impulsive Interference	56.65	0	15.35	0
	Δ = .5		Δ = 1			Δ = .5		Δ = 1	
m = 6	ASN	Errors	ASN	Errors	m = 6	ASN	Errors	ASN	Errors
No Impulsive Interference	40.25	0	10.65	0	No Impulsive Interference	40.45	0	10.65	0
Impulsive Interference	41.65	0	10.95	3	Impulsive Interference	48.38	0	14.63	0

(a) Normal partition space, $\alpha = \beta = .0067$

(b) Reduced partition space, $c_L = .0001$, $c_U = .9999$, $\alpha = \beta = .0067$

Note: Impulses occurred 20% of the time.

Joint Services Technical Advisory Committee
F44620-74-C-0056

R.F. Dwyer and L. Kurz

Naval Underwater Research Center
New London, Connecticut

REFERENCES

1. R.F. Dwyer and L. Kurz, "Relative Performance of Sequential and Fixed Sample Size Partition Detectors," Progress Report No. 41 to JSTAC, Polytechnic Inst. of New York, Report No. R-452.41-76, pp. 358-364 (November 1976).
2. R.F. Dwyer and L. Kurz, "Bivariate Dependent Sequential Partition Detectors," Progress Report No. 41 to JSTAC, Polytechnic Inst. of New York, Report No. R-452.41-76, pp. 353-357 (November 1976).
3. W.J. Bushnell and L. Kurz, "The Optimization and Performance of Detectors Based on Partition Tests," Proc. 1974 Allerton Conference on Circuits and Systems.
4. S.S. Rappaport and L. Kurz, "Optimal Decision Thresholds for Digital Signalling in Nongaussian Noise," 1965 IEEE International Convention Record, Part 2, pp. 198-212.

5. A. Wald, "Sequential Analysis," (New York: John Wiley and Sons, 1947).

6. M. Girshick, "Contributions to the Theory of Sequential Analysis," Annals of Math. Stat., Vol. 17, 1946.

7. J. Proakis, "Exact Distribution Functions of Test Length for Sequential Processors with Discrete Input Data," IEEE Trans. Inform. Theory, Vol. IT-9, No. 3, p. 182 (July 1963).

8. P. Kersten and L. Kurz, "M-Interval Detection of Nonconstant Signals," Progress Report No. 40 to JSTAC, Polytechnic Inst. of New York, Report No. R-452.40-75, pp. 367-370 (November 1975).

9. W.J. Bushnell and L. Kurz, "Selection of Near-Optimal Partitions in M-Interval Detectors," Progress Report No. 40 to JSTAC, Polytechnic Inst. of New York, Report No. R-452.40-75, pp. 348-356 (November 1975).

SCORE ESTIMATION FOR THE GENERALIZED QUANTILE DETECTOR

J. M. Habib and L. Kurz

The class of generalized quantile tests T_i^* was described in the previous report.[1] In this report, a recursive procedure for score estimation of the generalized quantile test is introduced which is based on the Kiefer-Wolfowitz procedure. In particular, the scores are selected recursively so that the efficacy of the detector is maximized. The procedure is easy to implement and operates on the data samples directly.

A. The Detector

Consider n_i i.i.d samples from the distribution F_i, $i = 1, \ldots, M$. Form the empirical distribution functions

$$F_{in_i}(x) = n_i^{-1} \sum_{j=1}^{n_i} u(x - x_{ji}) \tag{1}$$

where $u(.)$ is the unit step function (continuous from the right) and x_{ji} is the jth sample from F_i. Define the weighted sum of the empirical distributions

$$H_N(x) = \sum_{i=1}^{M} \beta_i F_{in_i}(x) \tag{2}$$

where

$$\sum_{1}^{M} \beta_i = 1, \ \beta_i \geq 0, \ \sum_{1}^{M} n_i = N \tag{3}$$

Then

$$T_i = \int_0^1 F_{in_i} H_N^{-1}(\alpha) \, d\nu(\alpha) \tag{4}$$

where $\nu(\alpha)$ is a bounded measure on 0, 1. Equation (4) corresponds to a general representation for a broad class of nonparametric statistics as shown in Reference 1. Let ν be the measure which assigns weight C_j to the points

$$0 < \lambda_1 < \ldots < \lambda_m < 1 \quad .$$

Define

$$T_i^* = \sum_{j=1}^{M} C_j W_i(\lambda_j) \qquad i = 1, \ldots, M \tag{5}$$

where

$$W_i(\lambda_j) = F_{in_i}[H_N^{-1}(\lambda_j)] \tag{6}$$

and

C_i are the scores of the test

M is the number of thresholds

It was also shown in Ref. [1] that the efficiency of the T_i^* test is high for small m, indicating that the penalty of using simple tests is small compared to the advantage gained in robustness.

The efficacy of a T test, ϵ (T) is defined by

$$\epsilon\ (T) = \left[\frac{\partial}{\partial \theta} E\ (T) \Big|_{\theta = 0} \right]^2 / \text{var}\ (T) \Big|_{\theta = 0} \tag{7}$$

where θ is a parameter defining separation between F_1 and F_2.

For the T_i test, Eq. (7) reduces to

$$\epsilon\ (T_i) = \frac{\vec{C}^T B \vec{C}}{\vec{C}^T \sum_o \vec{C}} \tag{8}$$

where

$$\vec{C} = (C_1 \ldots\ldots C_m) \tag{9}$$

is the vector of scores
and B, $\sum_o$ are mxm symmetric matrices with elements

$$b_{jk} = \frac{\partial W_i(\lambda_j)}{\partial \theta} \Big|_{\theta = 0} \cdot \frac{\partial W_i(\lambda_k)}{\partial \theta} \Big|_{\theta = 0} \tag{10}$$

$$\sigma_{jk} = \text{cov.}\ \{ W_i(\lambda_j),\ W_i(\lambda_k) \} \tag{11}$$

for $\theta = 0$, Eq. (11) reduces to

$$\sigma_{jk} = (\lambda_j \vee \lambda_k) [1 - (\lambda_j \wedge \lambda_k)] \tag{12}$$

where

$$(a \vee b) = \min \{ a, b \} \tag{13}$$

and

$$(a \wedge b) = \max \{ a, b \} \tag{14}$$

It was shown in Ref. [1], that under certain conditions, the scores which maximize Eq. (8) and the corresponding efficacy, may be calculated by forming a linear transformation:

$$\epsilon\,(T_1) = \vec{Y}_N^T \left(\sum^{-1/2} B \sum^{-1/2} \right) \vec{Y}_N \tag{15}$$

where

$$\vec{Y}_N = \sum^{-1/2} \vec{C}_N \tag{16}$$

and

$$\vec{Y}^T \vec{Y} = 1 \tag{17}$$

and that the efficacy is equal to the largest eigenvalue of $B \sum^{-1}$ and the optimum scores are the eigenvector corresponding to that eigenvalue .

B. A Recursive Approach to Score Estimation

In this section, it is proposed to estimate the optimum scores recursively. The efficacy function $\epsilon\,(\vec{C})$ is a convex function, with a maximum at $\vec{C}_{opt.}$. Noting that any score vector $\vec{C}$ corresponds to some p. d. f. f(.), we can see that

$$F(\vec{C}) = E\,[\,\epsilon\,|\,\vec{C}\,] \tag{18}$$

is a convex function with a maximum at $\vec{C}_{opt.}$.

The optimum scores $\vec{C}_{opt}$ corresponding to the maximum of Eq. (18), are estimated recursively using the generalization of Kiefer-Wolfowitz stochastic approximation method to the multivariate case. The iterative procedure to be used will be described next and the necessary notations will be introduced.

The m scores of the test, $C_1, C_2, \ldots, C_m$ will be denoted by the m-dimensional vector $\vec{C}$.

A unit setting of the score C_i and zero setting of the other m-1 scores will be denoted by the unit vector

$$\vec{e}_i \quad , \quad i = 1, 2, \ldots, m \tag{19}$$

The iterative adjustment will be made by making observations on the efficacy function $\epsilon\,(\vec{C})$ at various score vector value settings as follows:

Suppose that we are at the setting $\vec{C}_n$. We then make 2m observations, $\epsilon_n^1, \epsilon_n^2, \ldots, \epsilon_n^{2m}$

where

$\epsilon_n^1 = \epsilon\,(\vec{C}_n + \alpha_n \vec{e}_i)$ is an observation with score vector setting at

$$\vec{C} = \vec{C}_n + \alpha_n \vec{e}_i \tag{19a}$$

and $\epsilon_n^2 = \epsilon\ (\vec{C}_n - \alpha_n \vec{e}_i)$ is an observation with score vector setting at $\vec{C} = \vec{C}_n - \alpha_n \vec{e}_1$ ... .

$$\epsilon_n^{2i-1} = \epsilon\ (\vec{C}_i + \alpha_n \vec{e}_i)$$

$$\epsilon_n^{2i} = \epsilon\ (\vec{C}_i - \alpha_n \vec{e}_i) \tag{20}$$

We then form the m-dimensional vector $\vec{\epsilon}_n$, whose i^{th} component is given by

$$\epsilon_n^{2i-1} - \epsilon_n^{2i} = \epsilon\ (\vec{C}_n + \alpha_n \vec{e}_i) - \epsilon\ (\vec{C}_n - \alpha_n \vec{e}_i)$$

$$i = 1, \ldots, m \tag{21}$$

The next score setting in the iterative procedure is

$$\vec{C}_{n+1} = \vec{C}_n - \frac{\gamma_n}{\alpha_n} \vec{\epsilon}_n \tag{22}$$

where $\{ \gamma_n \}$ and $\{ \alpha_n \}$ are sequences of positive numbers satisfying

$$\text{(i)} \quad \lim_{n \to \infty} \alpha_n = 0$$

$$\text{(ii)} \quad \sum_1^\infty \gamma_n = \infty \tag{23}$$

$$\text{(iii)} \quad \sum_1^\infty \gamma_n \alpha_n < \infty$$

$$\text{(iv)} \quad \sum_1^\infty \gamma_n^2 / \alpha_n^2 < \infty$$

(for example $\gamma_n = n^{-1}$, $\alpha_n = n^{-1/3}$).

Under the condition that there is a unique maximum of $F(\vec{C})$, Blum[2] showed the convergence of $\vec{C}_n$ to $\vec{C}_{opt.}$ almost surely (a. s.) with the proper choice of α_n and γ_n, namely

$$P\left[\lim_{n \to \infty} \vec{C}_n = \vec{C}_{opt.} \right] = 1 \tag{24}$$

C. Numerical Results and Examples

To test the score estimation procedure outlined above, it was applied to the examples with known numerical results. These examples together with the results using the original and the developed method for score estimation will be considered next. The results using the developed method are calculated using a computer program to simulate the multidimensional Kiefer-Wolfowitz s.a. This program is included in Appendix D of Reference 3.

Example 1

For a Gaussian shift parameter and m=3, it was shown that

$\lambda_1 = .165$

$\lambda_2 = .5$

$\lambda_3 = .835$

The optimum score vector is given by

$$\vec{C}_{opt.} = [\ .605 \quad .518 \quad .605\ ]$$

and the corresponding maximum efficacy is

$$\epsilon_{max} = .883$$

The developed stochastic approximation method was applied as follows: Assume an arbitrary $\vec{C}$

$$\vec{C} = \vec{C}_{opt} + \vec{\Delta}$$

where

$\vec{\Delta}$ is an arbitrary deviation vector

$\vec{\Delta} = [\ \Delta_1\ \Delta_2\ \Delta_3\ ]$

Δ_1 and Δ_2 are chosen arbitrarily and Δ_3 is obtained from

$$\vec{C}\,\vec{C}^T = 1$$

The results obtained are as follow

for $\Delta_1 = -0.4$

$\Delta_2 = -0.4$

$\Delta_3 = .137$

and

$$\alpha_n = \frac{.1}{\sqrt{n}} \quad , \qquad \gamma_n = \frac{.2}{n}$$

The initial deviation vector

$$\vec{\Delta} = [\ -0.4 \quad -0.4 \quad .137\ ]$$

converged to

$$\vec{\Delta}_6 = [\ -0.1 \quad .08 \quad .02\ ]$$

in only six iterations and the efficacy corresponding to this score vector was given by

$$\epsilon = .8735$$

i.e., the efficacy deviation $\Delta\epsilon = 1.07\%$.

Example 2

For a Gaussian shift parameter and m=2, it is known that $\lambda_1 = .27$, $\lambda_2 = .73$. The optimum score vector is given by

$$\vec{C}_{opt.} = [.707 \quad .707]$$

and the corresponding maximum efficacy is

$$\epsilon_{max} = .810$$

The results of applying the stochastic approximation method as in Example 1 are:

$$\Delta_1 = -.4 \quad , \qquad \Delta_2 = .245$$

and

$$\alpha_n = \frac{.1}{\sqrt{n}} \quad , \qquad \gamma_n = \frac{.5}{n}$$

The initial deviation vector

$$\vec{\Delta} = [\ -.4 \quad .245\ ]$$

converged to

$$\vec{\Delta}_4 = [\ -.0049 \quad .0051]$$

in only four iterations and the corresponding efficacy deviation

$$\Delta\epsilon = .04\%$$

Example 3

For a Cauchy shift parameter, it is known that for m = 2 and

$$\lambda_1 = .445 \qquad \lambda_2 = .555$$

AD-A050 265 POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 20/3
PROGRESS REPORT NO. 42, 15 SEPTEMBER 1976 - 14 SEPTEMBER 1977, --ETC(U)
NOV 77 A A OLINER F44620-74-C-0056
UNCLASSIFIED POLY-MRI-452.42-77 NL

5 OF 7
AD
A050 265



The optimum score vector is given by

$$\vec{C}_{opt} = [\ .707 \quad .707]$$

and the corresponding maximum efficacy is

$$\epsilon_{max} = .429$$

The results of applying the stochastic approximation method as in Examples 1 and 2 are:

$$\Delta_1 = -.3 \quad , \qquad \Delta_2 = .206$$

and $\alpha_n = \dfrac{.1}{\sqrt{n}}$, $\gamma_n = \dfrac{.5}{n}$

The initial deviation vector

$$\vec{\Delta} = [\ -.3 \quad .206]$$

converged to

$$\vec{\Delta}_6 = [\ 0.00 \quad 0.00]$$

in six iterations and the efficacy deviation

$$\Delta\epsilon = 0\%$$

Example 4

For a Gaussian shift parameter, it is known that for m= 6, and

$$\vec{\lambda} = [\ .055 \quad .190 \quad .390 \quad .952 \quad .81 \quad .945]$$

The optimum score vector is given by

$$\vec{C}_{opt.} = [.501 \quad .371 \quad .333 \quad .333 \quad .371 \quad .501]$$

and the corresponding maximum efficacy is

$$\epsilon_{max} = .956$$

The results of applying the stochastic approximation method using the developed computer program in Appendix D, of Reference 3, are as follows

$$\vec{\Delta} = [\ -.501 \quad -.371 \quad -.333 \quad -.333 \quad -.371 \quad +.499\ ]$$

and

$$\alpha_n = \frac{.1}{n} \quad , \qquad \gamma_n = \frac{.2}{n}$$

The initial deviation vector $\vec{\Delta}$, converged to

$$\hat{\vec{\Delta}} = [\ .2\ \ -.04\ \ -.01\ \ -.04\ \ -.02\ \ -.07\ ]$$

in 1000 iterations and the efficacy deviation is

$$\Delta\epsilon = .004$$

Several other examples were solved and the procedure converged in all the cases tried. However, the number of iterations required for convergence increased as the number of estimated scores increased. A summary of the solutions for various examples is shown in Figs. 1 to 4 for various score numbers.

It should be noted that this procedure is inexpensive in terms of CPU time. The maximum CPU time used for estimating six scores simultaneously was 2.24 CPU sec. to attain an efficacy deviation of .4%.

Joint Services Technical Advisory Committee
F 44620-74-C-0056

I. M. Habib and L. Kurz

RCA Globcom

REFERENCES

1. J. I. Cochraine and L. Kurz, "A Generalized Quantile Detector," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76, pp. 315-321 (1976).
2. J. A. Blum, "Multidimensional Stochastic Approximation Methods," Annals of Mathematical Statistics, Vol. 25, No. 4 (1954).
3. I. M. Habib, "Recursive Estimation and Adaptive Detection in Nongaussian Noise Environments," Ph. D. Dissertation, Polytech. Inst. of New York, (June 1977).

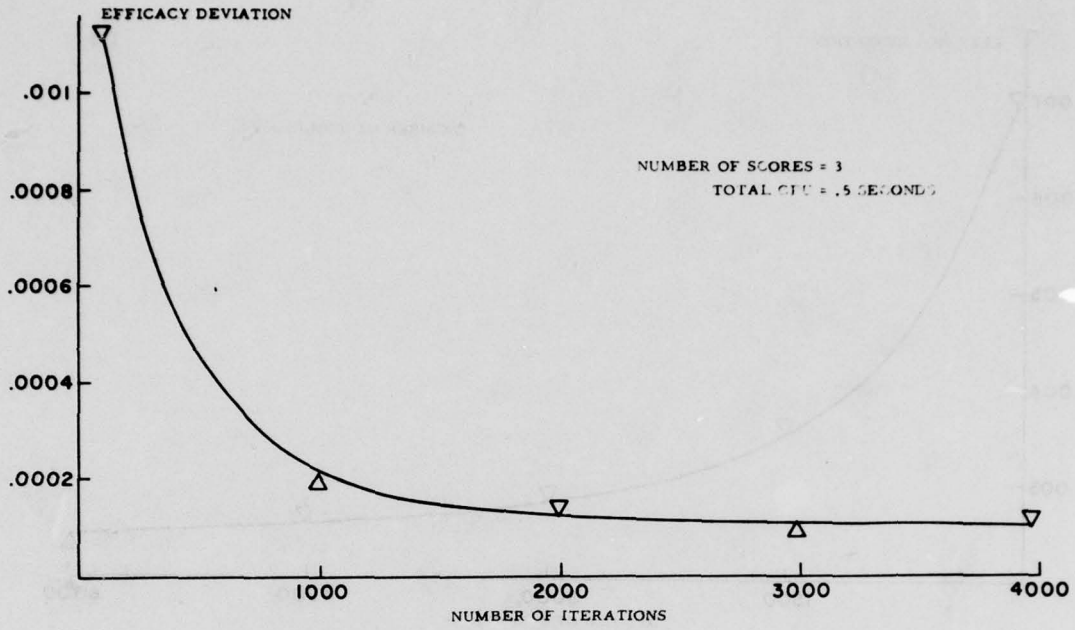


Fig. 1. Efficacy deviation of a generalized quantile test.

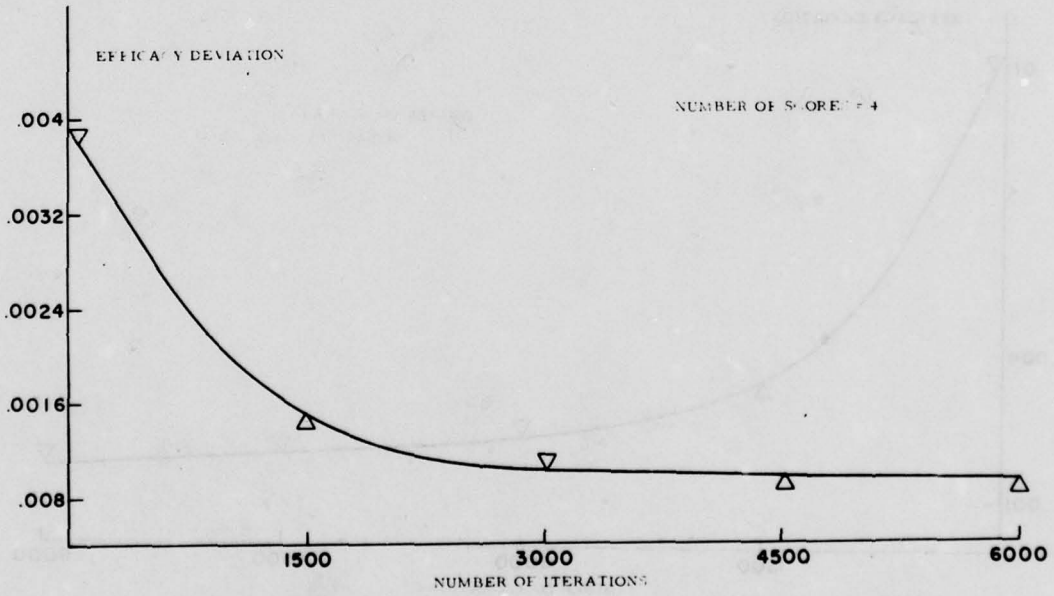


Fig. 2. Efficacy deviation of a generalized quantile test.

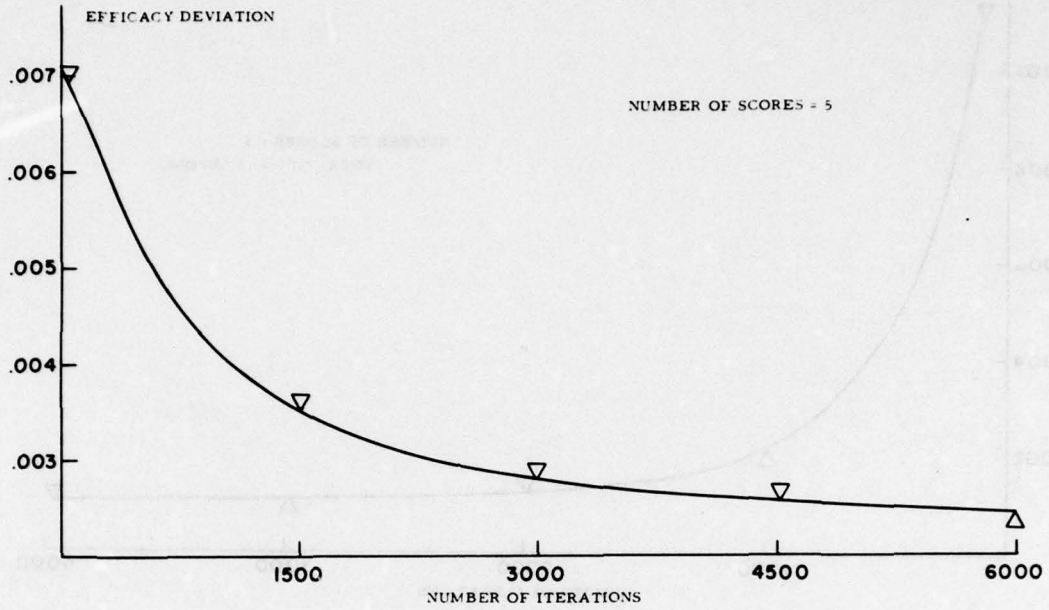


Fig. 3. Efficacy deviation of a generalized quantile test.

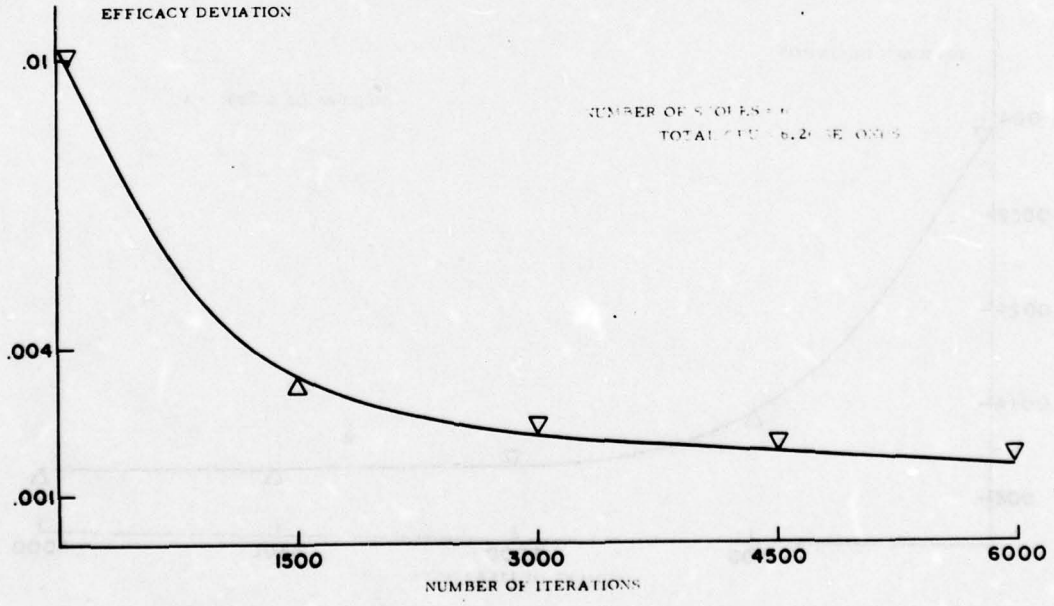


Fig. 4. Efficacy deviation of a generalized quantile test.

DATA SMOOTHING VIA ORDER STATISTICS

S-G. Tyan

Whenever a sequence of data is observed, be it an image or a speech parameter process, it is subject to "smoothing" at one or more of the many stages of signal processing. The purpose is to correct erratic data and to bring out the underlying signal in a clearer form.

Usually the underlying signal is thought to have a certain kind of "regularity." Thus, if the observed data sequence does not exhibit "irregular fluctuation" then smoothing is hardly necessary. In the past, with rare exceptions, people have used low-pass filters to carry out this task with notable success. However, in image signal processing, the story seems to be quite different. Here the edges are important information bearers and are usually abundant in many cases. Therefore, a low-pass filter which smears the jumps that always accompany the edges cannot accomplish the task it is supposed to do.

Instead of using the usual frequency contents of a signal as a measure of its "smoothness" or "regularity," we propose to use the supportedness[6] of a datum by its neighbors which is to be defined later. Before we introduce the upper envelope and the lower envelope of a data sequence, we shall offer a brief argument which goes like this: Let $\{x_n\}$ be a data sequence and let $\{\hat{x}_n\}$ be its smoothed version. Let k be a fixed positive integer. Suppose $\hat{x}_n$ is to be determined from its k successive neighbors and x_n itself. For example, we want to determine $\hat{x}_n$ from

$$x_{n-k+i}, \ldots, x_{n+1}$$

where i is an integer, $0 \le i \le k$. Since we lack any idea about the nature of the data sequence, we may assume that $\hat{x}_n$ lies in an interval bounded by the maximum and the minimum of $\{x_{n-k+i}, \ldots, x_{n+i}\}$, i.e.,

$$\min\{x_{n-k+i}, \ldots, x_{n+i}\} \le \hat{x}_n \le \max\{x_{n-k+i}, \ldots, x_{n+i}\}.$$

Since i can be any integer from 0 to k, thus

$$\max_{0 \le i \le k} \min\{x_{n+i-k}, \ldots, x_{n+i}\} \le \hat{x}_n \le \min_{0 \le i \le k} \max\{x_{n+i-k}, \ldots, x_{n+i}\}.$$

Definition: Let k be a fixed positive integer, and let $\{x_n\}$, $-\infty < n < \infty$, be a sequence. The two sequences $\{\overline{x}_n\}$ and $\{\underline{x}_n\}$ are the upper and lower envelopes of $\{x_n\}$, respeceively, where

$$\overline{x}_n = \min_{0 \le i \le k} \max\{x_{n+i-k}, \ldots, x_{n+i}\}$$

and

$$\underline{x}_n = \max_{0 \le i \le k} \min\{x_{n+i-k}, \ldots, x_{n+i}\} .$$

Remark: The $\{\overline{x}_n\}$ and $\{\underline{x}_n\}$ are simply the local minmax and the local maxmin.

Obviously $\{\underline{x}_n\} \le \{x_n\} \le \{\overline{x}_n\}$. Let I_n be the closed interval $[\underline{x}_n, \overline{x}_n]$. Then we shall say that x_n, the datum, is well-supported by its neighbors (with indices spanning from n-k to n+k) if $|\overline{x}_n - \underline{x}_n|$ is small and that it is not if $|\overline{x}_n - \underline{x}_n|$ is large. What is interesting about this definition is that if a datum x_n is supported by its neighbors on one side, say, $x_{n-k}, x_{n-k+1}, \ldots, x_{n-1}$, (i.e., they all lie in the proximity of x_n) then x_n according to the definition is well-supported by its neighbors. This property makes jump-type discontinuities almost harmless if one remembers how horrible they are in Fourier analysis. If $I_n = [\underline{x}_n, \overline{x}_n]$, which is sort of an eye, closes for all n, i.e., $\underline{x}_n = \overline{x}_n$, then we have the following theorem.

Theorem 1: The upper envelope and the lower envelope coincide with each other iff $\{x_n\}$ is a locally monotonic sequence of length k+2.

A sequence $\{x_n\}$ is locally monotonic of length m (abbr. LOMO(m)) iff $(x_n, x_{n+1}, \ldots, x_{n+m-1})$ is monotonic for all n.

If the upper and lower envelope of $\{x_n\}$ do not coincide or they are not even close to each other for all n, then the need arises of smoothing the data. It seems advisable to consider smoothers which belong to the $\mathcal{L}$ class.

Definition: A smoother T is said to belong to $\mathcal{L}$ iff the smoothed version, $T\{x_n\}$, is in between the upper and lower envelopes of $\{x_n\}$. In other words, we think that the undisturbed signal lies in between $\{\overline{x}_n\}$ and $\{\underline{x}_n\}$.

Theorem 2. $\mathcal{L}$ is closed under convex combination and compounding, i.e., if T_1 and T_2 are in $\mathcal{L}$, then $\alpha T_1 + (1-\alpha)T_2$, $0 \le \alpha \le 1$, and $T_1 * T_2$ are both in $\mathcal{L}$.

Examples of smoothers in $\mathcal{L}$ are numerous and some of them have been applied to speech processing[1] and image processing[2,3] with success. The most important of them are possibly the family of running medians.[4-6] A running median smoother of odd length 2p + 1 (abbr. RM(2p+1)) is defined as

$$RM(2p+1)\{x_n\} = \{y_n\} ,$$

where $y_n = \text{median } \{x_{n-p}, \ldots, x_{n+p}\}$ for all n. It is easy to check that RM(2p+1) belongs to $\mathcal{L}$ for all $0 \le p \le k$. From Theorem 1 and Theorem 2 we know that a LOMO(k+2) sequence is always invariant under any smoother made up of RM(2p+1), $0 \le p \le k$, through convex combination and compounding. In fact, it can be shown that

many of these smoothers have the power to drive an arbitrary sequence toward a LOMO(k+2) sequence while the jumps inherent in the signal sequence are preserved.

If we let $\overline{E}$ and $\underline{E}$ be two operators such that $\overline{E}\{x_n\} = \{\overline{x}_n\}$ and $\underline{E}\{x_n\} = \{\underline{x}_n\}$, respectively, then we have the following interesting theorem:

Theorem 3. Both $\underline{E}\overline{E}\{x_n\}$ and $\overline{E}\underline{E}\{x_n\}$ are locally monotonic sequences of length k+2. Furthermore: if $\{s_n\}$ is a LOMO(k+2) sequence which lies in between the two envelopes $\{\overline{x}_n\}$ and $\{\underline{x}_n\}$, then

$$\overline{E}\,\underline{E}\{x_n\} \le \{s_n\} \le \underline{E}\,\overline{E}\{x_n\} \quad .$$

Remark: $\underline{E}\,\overline{E}\{x_n\}$ is the largest LOMO(k+2) sequence which lies below $\{\overline{x}_n\}$ and $\overline{E}\underline{E}\{x_n\}$ is the smallest which is above $\{x_n\}$.

Theorem 4. $\overline{E}\,\underline{E}\{x_n\} \le \mathrm{RM}(2k+1)\{x_n\} \le \underline{E}\,\overline{E}\{x_n\}$.

Extensions of the above definitions and theories to arrays of data are currently under study. We are also examining the interplay between inverse filtering and nonlinear smoothing with application to image restoration.

Joint Services Technical Advisory Committee
F44620-74-C-0056

S-G. Tyan

REFERENCES

1. L. R. Rabiner, M. R. Sambur and C. E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," IEEE, Vol. ASSP-23, pp. 552-557 (1975).

2. G. W. Wecksung and K. Campbell, "Digital Image Processing at EG&G," IEEE Computer, Vol. 7, No. 5, pp. 63-71 (1974).

3. B. R. Frieden, "A New Restoring Algorithm for the Preferential Enhancement of Edge Gradients," J. Opt. Soc. Am., Vol. 66, No. 3, pp. 280-283 (1976).

4. J. W. Tukey, "Lecture Notes on Special Topics in Statistics," Statistics 411, (Princeton University, 1974).

5. A. E. Beaton and J. W. Tukey, "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data," Technometrics, Vol. 16, pp. 147-185 (1974).

6. P. F. Velleman, "Robust Non-Linear Data Smoothing," Technical Report No. 89, Series 2, Department of Statistics, (Princeton University, 1974).

THE NYQUIST PROBLEM IN THE PRESENCE OF NONLINEARITIES IN THE DUOBINARY SYSTEM

L. Kurz and M. Wernicki

In the previous work by the authors the effect of memoryless nonlinearities on intersymbol interference in a duobinary PAM system was considered.[1] It was demonstrated that these nonlinearities, natural to each system or introduced to suppress impulsive noise, create problems in controlling intersymbol interference. In a parallel research effort, a frequency domain approach to equalization in standard PAM systems in the presence of memoryless nonlinearity represented by a cubic polynomial was developed.[2] The main purpose of this report is to extend the results of Ref. 2 to duobinary systems.

A. Design in the Frequency Domain

Let the linear part of the signal have a spectrum $G_A(w)$ (desirable part), and the cubic term have a spectrum $G_B(w)$ (undesirable part). Using Nyquist's criterion, the constraint equations can be written as

$$\sum_n{}' G_A^{(n)}(w)\, G_R^{(n)}(w) = T x_o \qquad |w| \le \frac{\pi}{T} \tag{1}$$

and

$$\sum_n{}' G_B^{(n)}(w)\, G_R^{(n)}(w) = 0 \qquad |w| \le \frac{\pi}{T} \tag{2}$$

The minimization will be performed with respect to the variance of the noise (gaussian noise)

$$\sigma^2 = \frac{N_o}{2} \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} \sum_n{}' |G_R^{(n)}(w)|^2 \, dw \tag{3}$$

Thus, for each value of w, $|w| \le \frac{\pi}{T}$, one desires the stationary point of

$$J = \sum_n{}' \{G_R^{(n)}(w)|^2 + \lambda(w) G_A^{(n)}(w) G_R^{(n)}(w) + \gamma(w) G_R^{(n)}(w) G_B^{(n)}(w)\} \tag{4}$$

which is found using standard techniques of the calculus of variations, obtaining

$$G_R^{(n)}(w) = -\tfrac{1}{2}\lambda(w)[G_A^{(n)*}(w)] - \tfrac{1}{2}\gamma(w)[G_B^{(n)*}(w)] \tag{5}$$

where * denotes the complex conjugate operation.

Substituting Eq. (5) into Eqs. (1) and (2), yields

$$\sum_n \{-\tfrac{1}{2}\lambda(w)\left|G_A^{(n)}(w)\right|^2 - \tfrac{1}{2}\gamma(w)\left|G_A^{(n)}(w)\right|[G_B^{(n)}(w)]^*\} = Tx_o \tag{6}$$

$$\sum_n \{-\tfrac{1}{2}\lambda(w)[G_A^{(n)}(w)]^* G_B^{(n)}(w) - \tfrac{1}{2}\gamma(w)\left|G_B^{(n)}(w)\right|^2\} = 0 \tag{7}$$

Solving Eq. (7) for $\gamma(w)$ results in

$$\gamma(w) = \frac{-\sum_n \lambda(w)[G_A^{(n)}(w)]^* G_B^{(n)}(w)}{\sum_n \left|G_B^{(n)}(w)\right|^2} \tag{8}$$

Using Eqs. (8) and (6), and solving for $\lambda(w)$, yields

$$\lambda(w) = \frac{Tx_o}{\sum_n \{\frac{-\left|G_A^{(n)}(w)\right|^2}{2} + [\frac{\sum_i G_A^{(i)}(w)G_B^{(i)}(w)}{2\sum_i \left|G_B^{(i)}(w)\right|^2}]\, G_A^{(n)}(w)[G_B^{(n)}(w)]^*\}} \tag{9}$$

Similarly,

$$\gamma(w) = \frac{Tx_o}{\sum_n \{[\frac{\sum_i \left|G_B^{(i)}(w)\right|^2}{2\sum_i [G_A^{(i)}(w)]^* G_B^{(i)}(w)}]\, \left|G_A^{(n)}(w)\right|^2 - \frac{G_A^{(n)}(w)[G_B^{(n)}(w)]^*}{2}\}} \tag{10}$$

The desired filter characteristic is obtained by substituting Eqs. (9) and (10) into Eq. (5), obtaining

$$G_R^{(n)}(w) = \frac{Tx_o[G_B^{(n)}(w)]^*}{\sum_n \{G_A^{(n)}(w)[G_B^{(n)}(w)]^* - [\frac{\sum_i \left|G_B^{(i)}(w)\right|^2}{\sum_i [G_A^{(i)}(w)]^* G_B^{(i)}(w)}]\, \left|G_A^{(n)}(w)\right|^2\}} +$$

$$+ \frac{T x_o [G_A^{(n)}(w)]^*}{\sum_n \{|G_A^{(n)}(w)|^2 - [\frac{\sum_i [G_A^{(i)}(w)]^* G_B^{(i)}(w)}{\sum_i |G_A^{(i)}(w)|^2}] G_A^{(n)}(w)[G_B^{(n)}(w)]^*\}} \quad (11)$$

The number of terms for which intersymbol interference can be eliminated depends upon the available bandwidth. The cubic term produces power terms and cross-product terms with three times the original signal bandwidth. This extra bandwidth introduces the additional degrees of freedom in the receiver design, allowing the additional constraints to be satisfied.

B. Application of the Design Procedure to Duobinary Signals

An example will help to clarify the use of this filter design procedure. Specifically, consider the duobinary signal given by

$$S(w) = 2T \cos\left(\frac{T}{2}w\right) \qquad |w| \leq \frac{\pi}{T}$$

$$= 0 \qquad \text{otherwise}$$

It may be shown that such signal satisfies Nyquist's criterion for eliminating the intersymbol interference. However, as with an ordinary PAM system, it is advantageous to split the overall duobinary signal shaping equally between the transmitter and receiver.

The frequency response due to the linear portion of nonlinearity is shown in Fig. 1, curve A. The cubic term of the nonlinearity was calculated using IBM 370 computer and its frequency response is shown in Fig. 1, curve B. The frequency response corresponding to Eq. (11) is shown in Fig. 1, curve C.

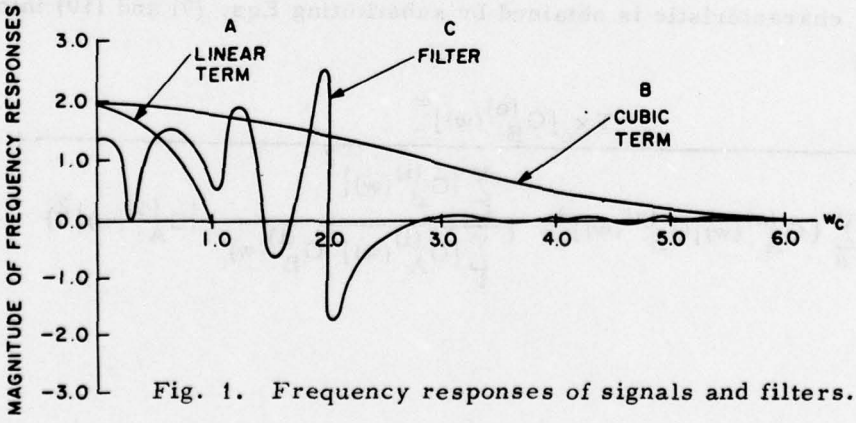


Fig. 1. Frequency responses of signals and filters.

The same constraints as in Ref. 2, an equalizer (filter) with transfer characteristics satisfying these constraints may be designed. The time response of this equalizer, as shown in Fig. 2, is not well damped, increasing the jitter problem. The frequency response, shown in Fig. 1, reveals poor suppression properties of the equalizer.

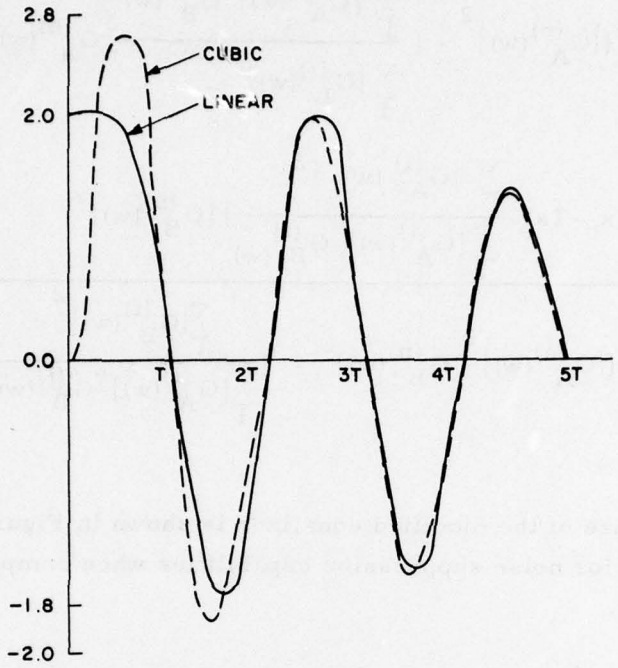


Fig. 2. Output time responses due to linear and nonlinear terms.

Faced with these difficulties, it is desirable to seek a better equalizer which may allow small intersymbol interference due to the linear and cubic terms yet, at the same time, reduce the jitter and noise problem. This approach is preferable because it results in better overall system performance.

To this end, the following changes need be made in the constraint Eqs. (1) and (2):

$$\sum_n G_A^{(n)}(w)G_R^{(n)}(w) = T x_o \ , \qquad |w| \leq \frac{\pi}{T} \tag{12}$$

$$\sum_n G_B^{(n)}(w)G_R^{(n)}(w) = T x_o \qquad |w| \leq \frac{\pi}{T} \tag{13}$$

Following the procedure leading to Eq. (11), one obtains the transfer characteristic of a modified equalizer in the form

$$G_{R_{comp}}^{(n)}(w) = \frac{\{Tx_o - Tx'_o \frac{\sum_i G_A^{(i)}(w)[G_B^{(i)}(w)]^*}{\sum_i |G_B^{(i)}(w)|^2}\}[G_A^{(n)}(w)]^*}{\sum_n \{|G_A^{(n)}(w)|^2 - [\frac{\sum_i [G_A^{(i)}(w)]^* G_B^{(i)}(w)}{\sum_i |G_B^{(i)}(w)|^2}] G_A^{(n)}(w)[G_B^{(n)}(w)]^*\}} +$$

$$+ \frac{\{Tx_o - Tx'_o \frac{\sum_i |G_A^{(i)}(w)|^2}{\sum_i [G_A^{(i)}(w)]^* G_B^{(i)}(w)}\}[G_B^{(n)}(w)]^*}{\sum_n \{|G_A^{(n)}(w)|^2 [G_B^{(n)}(w)]^* - [\frac{\sum_i |G_B^{(i)}(w)|^2}{\sum_i [G_A^{(i)}(w)]^* G_B^{(i)}(w)}] |G_A^{(n)}(w)|^2\}} \qquad (14)$$

The frequency response of the modified equalizer is shown in Figure 3. The new equalizer has a superior noise suppression capabilities when compared with the equalizer of Figure 1.

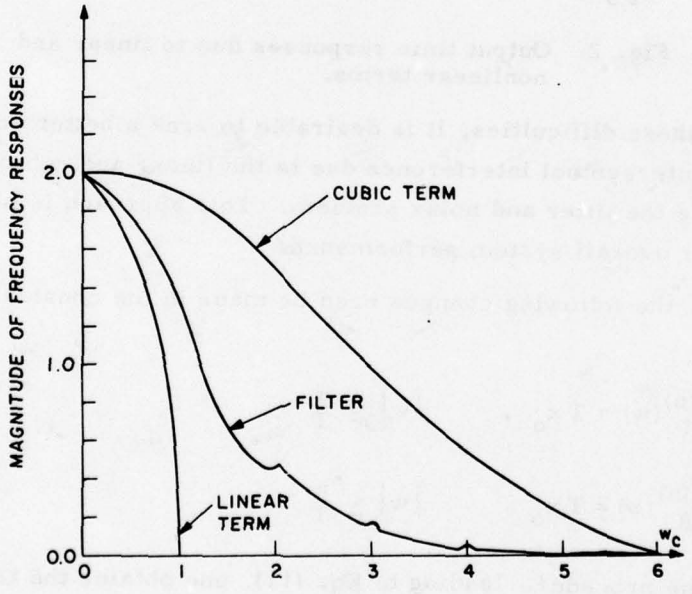


Fig. 3. Frequency response of the filter with inter-symbol term remaining.

Since the time response of the new equalizer is well damped (see Fig. 4), it has good jitter immunity. The small disadvantage of a tolerable intersymbol interference which is a function of the nonlinearity, the noise, and the signal amplitude, is a small price to pay for an overall improvement in system performance.

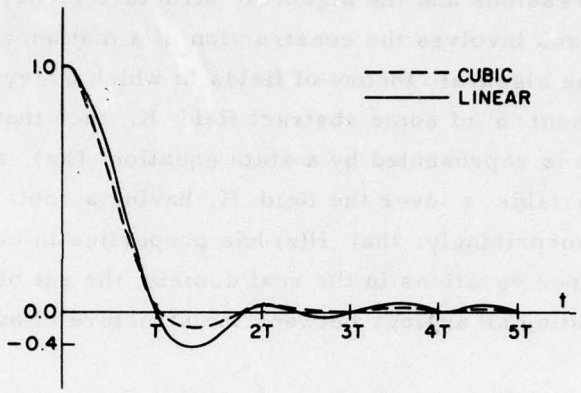


Fig. 4. Output time responses for filter with intersymbol interference remaining.

Joint Services Technical Advisory Committee
F44620-74-C-0056

L. Kurz and M. Wernicki

REFERENCES

1. L. Kurz and M. Wernicki, "Deterioration in Performance of a Duobinary PAM System by Introduction of Nonlinearities," Progress Report No. 41 to JSTAC, Polytechnic Institute of New York, Report No. R-452.41-76, pp. 348-352 (1976).

2. L. Kurz and G. Soloway, "The Nyquist Problem in the Presence of Nonlinearities in the Data Transmission System," Progress Report No. 41 to JSTAC, Polytechnic Institute of New York, Report No. R-452.41-76, pp. 322-327 (1976).

A NEW APPROACH TO THE ALGEBRAIC THEORY OF AUTOMATA

T. T. Lee and E. J. Smith

The objective of this research is to develop a new approach to the theory of automata which will provide the formalism necessary to encompass the concepts of the algebra of regular expressions and the algebraic structure theory of automata. More specifically, the approach involves the construction of a mathematical model of automata based upon the algebraic theory of fields in which every regular set R is represented by an element α of some abstract field K, such that the automation which accepts the regular set is represented by a state equation H(z), a certain polynomial in the indeterminate variable z over the field K, having a root α. Preliminary work has revealed perhaps surprisingly, that H(z) has properties in common with those of differential and difference equations in the real domain; the set of roots of H(z) forms an affine space, suggesting an analogy between the structure of automata and that of linear systems.

In the following sections of this report, the preliminary work and results obtained are presented and the proposed future work is outlined.

A. Basic Ideas and Preliminary Work

We begin by introducing some symbols and terminology which for the most part follow standard usage.

(1) $\sum = \{0,1\}$, the input alphabet of an automaton having two input symbols

(2) $\sum^* = \{\lambda, 0, 1, 00, 01, 10, \ldots\}$, the set of words generated by $\sum$ under concatentation

(3) $\mu: \sum^* \to N$ a bijective mapping from $\sum^*$ into N, the set of nonnegative integers defined recursively by

(i) $\mu(\lambda) = 0$

(ii) If $\mu(\omega) = n$, then

$\mu(\omega\ 0) = 2n+1$ and $\mu(\omega\ 1) = 2$ for all $\omega \in \sum^*$

(4) F, the prime field of characteristic 2, i.e., GF (2)

(5) $F[x] = \{f(x) \mid f(x) = \sum_{n=-\infty}^{\infty} a_n x^n,\ a_n \in F\}$,

the extension of the formal power series in the indeterminate variable x over F

(6) $F[x]^+ = \{g(x) \mid g(x) = \sum_{n=0}^{\infty} b_n x^n,\ b_n \in F\}$,

the ring of the formal power serves (polynomial) in x over F

(7) $\hat{\sum}^*$, the power set of $\sum^*$.

We now define a mapping which provides the link between a regular set R, and a formal power series in x.

$$\Phi : \hat{\sum}^* \rightarrow F[x]$$

For any subset $L \in \hat{\sum}^*$, there is an associated formal power series $f_L(x)$ defined by

$$\Phi(L) = f_L(x) = \sum_{\omega \in L} x^{\mu(\omega)}$$

Conversely, for any power series

$$f(x) = \sum_{n=0}^{\infty} a_n x^n$$

there is a subset $R \in \hat{\sum}^*$, such that $\omega \in R$ iff $a_{\mu(\omega)} = 1$.

The mapping, Φ, is bijective and posseses the following properties:

(i) $\Phi(\sum^*) = f_{\sum^*}(x) = 1 + x + x^2 + \ldots = \frac{1}{1+x}$

(ii) $\Phi(A \cup B) = \Phi(A) + \phi(B)$, provided that $A \cap B = \phi$,

i.e., $f_{A \cup B}(x) = f_A(x) + f_B(x)$

(iii) $\Phi(A \cap B) = \Phi(A) \odot \Phi(B)$, where

$\Phi(A) \odot \Phi(B)$ denotes the Hadamand product of $\Phi(A)$ and $\Phi(B)$

i.e., if $f_A(x) = \sum a_n x^n$ and $f_B(x) = \sum b_n x^n$, then

$f_{A \cap B} = \sum a_n b_n x^n$

(iv) $\Phi(\phi) = 0$, where ϕ is the empty set

(v) $\Phi(\overline{L}) = \Phi(\sum^*) - \Phi(L) = \frac{1}{1+x} + \Phi(L)$,

where $\overline{L}$ is the complement of L.

Considering $F[x]$ as a vector space over F, we now introduce a linear operator, D,

$$D: F[x] \to F[x],$$

such that $Df(x) = f^2(x)$ for all $f(x) \in F[x]$, where $f^2(x) = [f(x)]^2$. It follows that, given $R \subseteq \sum^*$, $\Phi(R) = f_R(x) \in F[x]$, $R0 = \{\omega 0 \mid \omega \in R\}$, and $R1 = \{\omega 1 \mid \omega \in R\}$, then

$$\Phi(R0) = \sum_{\omega \in R} x^{2\mu(\omega)+1} = xDf_R = xf_R^2$$

The use of the operator, D, will be illustrated in later examples. However, we first note for reference some terminology describing several special classes of polynomials.

(1) A polynomial $L(z)$ over $F[x]$ is called a linearized polynomial iff

$$L(z) = \sum_{i=0}^{n} a_i z^{2^i} = \left(\sum_{i=0}^{n} a_i D^i\right) z$$

A polynomial $L(z)$ is a linearized polynomial iff its roots form a vector space over F.

(2) A polynomial $A(z)$ over $F[x]$ is called an affine polynomial iff

$$A(z) = L(z) + \xi$$

where $L(z)$ is a linearized polynomial and $\xi \in F[x]$.

A polynomial $A(z)$ is an affine polynomial iff its roots form an affine space over F, i.e., if $a_1, a_2, \ldots, a_n$ are roots, then $a_1 + \sum c_k(a_k - a_1)$ is also a root for all $c_k \in F$.

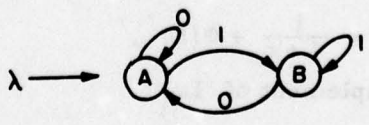
The following analogies are suggested:

(i) Linearized Polynomial $\longleftrightarrow$ Homogeneous Differential Equation

(ii) Affine Polynomial $\longleftrightarrow$ Non-Homogeneous Differential Equation.

B. Applications to Finite Automata

We now illustrate by examples the application of some of the previous concepts to finite state machines.



Let R_1 and R_2 be the sets of words (i.e., strings) accepted by states A and B, respectively, of the given automaton. Then,

$$R_1 = R_1 0 + R_a 0 + \lambda$$

$$R_2 = R_1 1 + R_2 1$$

Our objective is to find the formal power series (i.e., polynomial) representations corresponding to R_1 and R_2. Let $\Phi(R_1) = f$ and $\Phi(R_2) = g$; then, using the D-operator notation,

$$f = xDf + xDg + 1$$

$$g = x^2 Df + x^2 Dg$$

or in matrix form

$$\begin{bmatrix} f \\ g \end{bmatrix} = \begin{bmatrix} x & x \\ x^2 & x^2 \end{bmatrix} \cdot \begin{bmatrix} Df \\ Dg \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

It follows that

$$f = xf^2 + xg^2 + 1$$

$$g = x^2 f^2 + x^2 g^2$$

By simple algebra, we find that f and g satisfy the affine polynomials, H(f) and H(g), respectively.

$$(x^3 + x) f^2 + f + (x^3 + 1) = \longleftrightarrow H(f) = 0$$

$$(x^2 + 1) g^2 + g + x^2 = \longleftrightarrow H(g) = 0$$

After rearragement, we have the factored forms,

$$\{(x^2 + x) f + (x^2 + x + 1)\} \cdot \{xf + (x + 1)\} = 0$$

$$\{(x^2 + 1) g + x^2\} \cdot \{g + 1\} = 0 .$$

Solving for f and g,

$$f = \frac{x^2 + x + 1}{x^2 + 1} = 1 + \frac{x}{1 + x^2} , \frac{x + 1}{x}$$

$$g = \frac{x^2}{1 + x^2} , 1$$

We reject the second roots for f and g, since,

$$f \neq \frac{x+1}{x} = 1 + x^{-1}, \text{ i.e., } x^{-1} \notin F[x]^+$$

and

$$g \neq 1, \quad \lambda \notin R_2 .$$

Thus for the given automaton we have characterized the regular expressions R_1 and R_2 by algebraic expressions f and g, respectively. To illustrate further, let us begin with R_1 and R_2 which can be found by conventional methods to be

$$R_1 = \lambda \cup \sum^* 1 .$$

We know from previously that $\Phi(\sum^*) = \frac{1}{1+x}$.

Therefore,

$$\Phi(R_1) = f = 1 + xD\,\Phi(\sum^*) = 1 + xD\,\frac{1}{1+x}$$

$$= 1 + \frac{1}{1+x^2}$$

$$\Phi(R_2) = f = x^2 D\Phi(\sum^*) = x^2 D\,\frac{1}{1+x} = \frac{x^2}{1+x^2}$$

which checks our previous solution for f and g. Thus for any deterministic automaton we can construct "state equations," i.e., H(f) and H(g), whose roots are the formal power series corresponding to regular expressions. The analogy with linear dynamic systems is evident.

The affine polynomial corresponding to an automaton may or may not be factorable; in the latter case, the formal power series representation, f(x) of the corresponding regular set, R, cannot be represented as a rational fraction of the form p(x)/g(x), where p(x) and g(x) are polynomials in x. Thus, a given regular set may or may not be describable by a closed-form algebraic expression in x. It is conjectured that a formal power series f(x) which is describable by a rational fraction of the form p(x)/g(x) represents a regular set R, under the mapping, $\Phi(R)$.

There is the need for the development of additional techniques for manipulating formal power series expressions. We can generalize the previous equations for $\Phi R0) = xD\Phi(R)$ and $\Phi(R1) = x^2 D\Phi(R)$ as follows: $\Phi(R\omega) = x^{H(\omega)} D^{|\lambda|}\,\Phi(R)$, where $|\omega|$ denoted the length of ω .

Example 2

$$\Phi(010) = xD\Phi(R01) = xDx^2D\Phi(R0)$$

$$= xDx^2DxD\,\Phi(R) = x^5D^2xD\,\Phi(R)$$

$$= x^9D^3\,\Phi(R)$$

where $\mu(010) = 9$ and $|010| = 3$

The result follows directly from the formula for $\Phi(R\omega)$.

The formal power series of a given regular set satisfies some affine polynomial, and the polynomial can be constructed from the state diagram of the corresponding automaton. Given the regular expression for some regular set, one can construct the affine polynomial for the corresponding power series. The following two examples provide further illustrations of the application of the D-operator.

Example 3

Given $R = 1^*$, let $\Phi(R) = f(x)$.

Then, $f(x) = 1 + x^2D1 + x^2Dx^2D1 + \ldots$

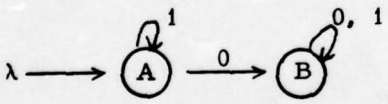
$$= 1 + x^2D + (x^2D)^2 \ldots$$

$$= \frac{1}{1 + x^2D}$$

Hence, $(1 + x^2D)\, f = 1$

Therefore, $x^2f^2 + f + 1 = 0$

The state diagram of the automaton which accepts 1^* is given below.



Starting with the implicit equation for R_A.

$$R_A = R_A\,1 + \lambda$$

and letting $\Phi(R_A) = f(x)$, we obtain.

$$f = x^2Df + 1 = x^2f^2 + 1$$

Hence,

$$x^2f^2 + f + 1 = 0$$

which agrees with the previous result.

In this example, the affine polynomial is irreducible and hence f(x) cannot be given in closed form.

Example 4

Consider the identity, $(01)^+ = 0(10)^* 1$. First we let $R = (01)^+$ and $\Phi(R) = f(x)$. For $f = \Phi(01)^+$), $\mu(01) = 4$ and $|01| = 2$.

Then, $f(x) = x^4 + (x^4D^2)x^4 + (x^4D^2)^2x^4 + \ldots$

$$= \frac{1}{1 + x^4D^2} \cdot x^4$$

which yields, $f + x^4D^2f = x^4$

or $x^4f^4 + f + x^4 = 0$

Alternatively, let $R = 0(10)^* 1$ and $\Phi(R) = f(x)$. Now $\mu(10) = 5$ and $|10| = 2$. Proceeding algebraically as before, one obtains,

$$f = x^2D(1 + x^5D^2)^{01} x$$

Continuing,

$$D^{-1}x^{-2} = (1 + x^5D^2)^{-1} x$$

$$(1 + x^5D^2)D^{-1}x^{-2}f = x$$

$$D^{-1}x^{-2}f + x^5D\,x^{-2}f = x$$

$$D^{-1}x^{-2}f + xf^2 = x$$

$$x^{-2}f + dxf^2 = Dx$$

$$x^{-2} + x^2f^4 = x^2$$

Finally, $x^4f^4 + f + x^4 = 0$

The same result obtains as before. Equivalent regular expressions define the same affine polynomial

Formal rules for the manipulation of operator polynomials have yet to be developed.

C. Concepts Related to Structural Properties of Automata

In the remainder of this section we illustrate some preliminary ideas which are related to the structural properties of automata.

Example 5

Consider the six-state finite automaton, L, described by the state table and set of regular equations shown below.

L	0	1
1	2	3
2	1	3
3	4	5
4	3	2
5	1	6
6	1	5

$$R_1 = R_2 0 + R_5 0 + R_6 0$$
$$R_2 + R_1 0 + R_4 1$$
$$R_3 + R_1 1 + R_2 1 + R_4 0$$
$$R_4 + R_3 0$$
$$R_5 = R_3 1 + R_6 1$$
$$R_6 = R_5 1$$

Let $Z_i = \Phi(R_i) \in F[x]$ for $i = 1, 2, \ldots, 6$

Then, from previous work,

$$\begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Z_5 \\ Z_6 \end{bmatrix} = \begin{bmatrix} 0 & X & 0 & 0 & X & X \\ X & 0 & 0 & X^2 & 0 & 0 \\ X^2 & X^2 & 0 & X & 0 & 0 \\ 0 & 0 & X & 0 & 0 & 0 \\ 0 & 0 & X^2 & 0 & 0 & X^2 \\ 0 & 0 & 0 & 0 & X^2 & 0 \end{bmatrix} \cdot \begin{bmatrix} DZ_1 \\ DZ_2 \\ DZ_3 \\ DZ_4 \\ DZ_5 \\ DZ_6 \end{bmatrix}$$

In general, the machine is characterized by the algebraic expression

$$Z = A\,D\,Z\,,$$

where $Z \in F[x]^6$, $F[x]^6$ denoted the six-fold Cartesian product of $F[x]$, and A is a 6×6 matrix over $F[x]$. The characteristic function of matrix A is readily f found to be

$$f_A(\lambda) = [\lambda + x][\lambda + x^2][\lambda + (x+x^2)][\lambda^3 + (x^2+x^3)\lambda + (x^4+x^5)]$$

It has been found that if π is a partition with the substitution property on the set of states of machine L and L_π is the image machine with the state equation $Z' = A'D\,Z'$, and $f_{A'}(\lambda)$ is the characteristic function of matrix A', then $f_{A'}(\lambda)$ is a divisor of $f_A(\lambda)$.

For the given machine, L, we observe that the following partition posses the substitution property

$$\pi = \{\overline{1,2}\,,\ \overline{3};\ \overline{4};\ \overline{5,6}\}$$

and, accordingly, that L_π given below is an image machine.

L_π	0	1
q_1	q_1	q_2
q_2	q_3	q_4
q_3	q_2	q_1
q_4	q_1	q_4

where

$q_1 = \overline{1,2}$

$q_2 = \overline{3}$

$q_3 = \overline{4}$

$q_4 = \overline{5,6}$

The matrix A' associated with machine L_π is found to be

$$A' = \begin{bmatrix} X & 0 & X^2 & X \\ X^2 & 0 & X & 0 \\ 0 & X & 0 & 0 \\ 0 & X^2 & 0 & X^2 \end{bmatrix}$$

and the characteristic function of A' is given by

$$f_{A'}(\lambda) = [\lambda + (x + x^2)][\lambda^3 + (x^2 + x^3) + (x^4 + x^5)]$$

which, by inspection, is seen to be a factor of $f_A(\lambda)$. If $f_A(\lambda)$ for a given machine L is irreducible, the machine does not possess an image machine. Thus we have a link between a structural property of a finite automaton and our algebraic characterization of the machine. A question to be explored further is the following: if $f_{A'}(\lambda)$ is a divisor of $f_A(\lambda)$, does this imply that the state space of machine L_π is an invariant subspace of the machine L? The relationship between the algebraic representation and other structural properties will be explored as well.

Another approach will be based upon the introduction of a certain linear transformation. For each partition π on $S = \{S_1, S_2, \ldots, S_n\} = \{S_i\}$, the set of states of machine L, there is a representation function,[8]

$r : S \longrightarrow S,$

such that,

$$S_i \equiv S_j(\pi) \quad \text{iff} \quad r(i) = r(j)$$

where the function r satisfies the following criteria:

(1) $r(i) \leq i$, for all i

(2) $r \cdot r(i) = r(i)$, for all i.

We now define the $n \times n$ matrix, T, induced by the representation function, r, such that

$$T = (p_{ij}) \quad , \quad p_{ij} \begin{cases} 1, \text{ if } f(j) = i \\ 0, \text{ if } f(j) \neq i \end{cases}$$

We can show that $T^2 = T$, as follows:

suppose $T^2 = (m_{ij}) = \left(\sum_{k=1}^{n} p_{ik} p_{kj} \right)$, we must demonstrate that

$$m_{ij} = \sum_{k=1}^{n} p_{ik} p_{kj} = p_{ij} \cdot$$

For each column of T, there is one and only one element equal to 1; all other elements equal 0. Therefore, if $m_{ij} = 1$, there is some ℓ for $1 \leq \ell \leq n$, such that $p_{i\ell} = p_{\ell j} = 1$. Then by definition, $r(\ell) = i$ and $r(j) = \ell$, which implies $r \cdot r(j) = r(\ell) = i$. By the idempotent property of r, $r(j) = i$ and thus $p_{ij} = 1$. Conversly, if $p_{ij} = 1$, then $r(j) = i$; then, $r(i) = r \cdot r(j) = r(j) = i$ and $p_{ii} = 1$. It follows that $\sum_{k=1}^{n} p_{ik} p_{kj} = p_{ii} p_{ij} = p_{ij} = m_{ij}$. Therefore, $T^2 = T$.

The transformation matrix T induced by previously given SP partition π on the set of states of the machine in Example 5 is as follows:

$$T = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

It is easy to verify that $T^2 = T$.

We know[9] that a linear transformation, T, on a vector space V is a projection iff $T^2 = T$. It follows that the transformation, T, induced by the partition π having the substitution is a projection over the state space $F[x]^n$. The properties possessed by such a transformation will be investigated. In general, the use of the T matrix as a tool for exploring structural properties of machines will be pursued.

D. Proposed Future Research

In accordance with the general approach for the development of a unified algebraic theory of automata outlined in the previous section, the proposed program will be pursued with the following objectives:

(1) To obtain a well-defined, unique representation of a regular set as an element of an abstract field K

(2) To develop procedures for setting up the state equation over K for a given automaton, such that the solution of the state equation is the representation of the regular set accepted by the automaton

(3) To analyze the properties of the affine space formed by the roots of the state equation of the automaton

(4) To investigate the structure and general criteria for decomposition of an automaton from the state equation associated with it

(5) To investigate the analogies between automata and linear systems, based upon the abstract field representation of the former.

In general, the goal is to develop a comprehensive algebraic theory of automata based upon the algebraic theory of fields.

Joint Services Technical Advisory Committee
F44620-74-C-0056

National Science Foundation
ENG73-04232

T. T. Lee and E. J. Smith

REFERENCES

1. B. L. Van der Waerden, "Modern Algebra," (New York: Frederick Unger, 1953).

2. A. N. Kolomogorov, "Foundations of Probability," (New York: Chelsea, 1956).

3. E. R. Berlekamp, "Albegraic Coding Theory," (New York: McGraw Hill, 1968), Chap. 11.

4. S. Eilenberg, "Automata, Languages, and Machines," Vol. A, (New York: Academic Press, 1974).

5. O. Ore, "On A Special Class of Polynomials," American Mathematical Society, Transactions, No. 35, pp. 559-614 (1933).

6. O. Ore, "Contribution to the Theory of Finite Fields," American Mathematical Society, Transactions, No. 36, pp. 234-274 (1934).

7. O. Ore, "Theory of Non-Commutative Polynomials," Annals of Mathematics, pp. 480-508 (1933).

8. T. Lee and E. J. Smith, "On the Partitioning Representation for Finite Sets and Its Aplication," Report POLY EE/EP-76-009, Polytech. Inst. of New York (May 1976).

9. P. R. Halmos, "Finite-Dimensional Vector Spaces," (D. van Nostrand, 1958).

A NEW COMPUTER HARDWARE DESCRIPTION LANGUAGE FOR THE STUDY OF COMPUTER STRUCTURES

M. Adamowicz and J. Mirza

A. Introduction

A new hardware description language referred to as MDSL (Microcomputer Design and Simulation Language) is described and its application to an illustrative example consisting of an array of identical processors operating asynchronously is shown. The advantages of such a language are severalfold: (a) it serves as a tool for studying the performance of various computer architectures, (b) it provides a well-defined notation for describing a particular architecture, (c) it permits the quantitative comparison of alternative system designs by writing test programs to be run on simulated machines, (d) it enables designers to investigate alternative designs and to evaluate various trade-offs before actually constructing the computer.

MDSL as a particular hardware description language was designed to be a highly structured, modular, register transfer type language that incorporates the control-structure characteristic of a microprogrammable system, such as the one proposed by Wilkes[1] in the early 1950's. As a consequence of the modular structure of MDSL and the decision to write it in PL/I, it was possible to implement quickly a basic subset of the full language for immediate use and to then gradually incorporate the more advanced features. The current version of MDSL (Version 2.00) allows, among other things, the capability of working with multiple copies of processors connected together in a network and with various types of content-addressable parallel processors.

A partial description of the language and its options is given in the present report. A more complete guide to the MDSL language is provided in Refs. 2 and 3, along with a number of examples showing its use in modelling various computers and computer-like forms.

B. MDSL Overview

The five basic components of the classical digital computer: input, output, memory, arithmetic and logic, and control are represented in Figure 1. These components communicate by means of signals that contain data, instructions, or control information. The control unit controls the sequencing and direction of these signal flows and therefore determines how the computer operates. Wilkes[1] proposed a model of microprogramming in order to systematize computer description by providing an orderly approach to the description of the control unit. Figure 2 contains a simplified representation of the Wilkes model. In this figure the horizontal lines are called microinstructions; the four major components of the model are as follows:

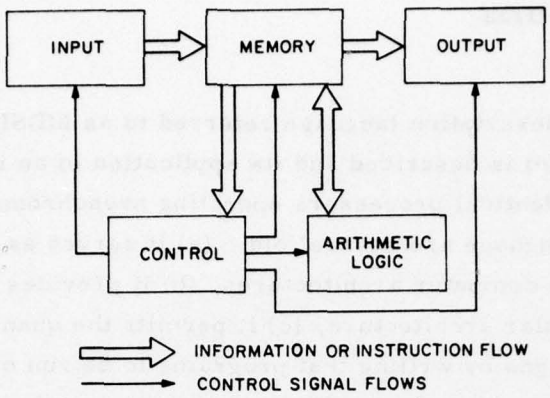


Fig. 1. Functional components of a modern digital computer.

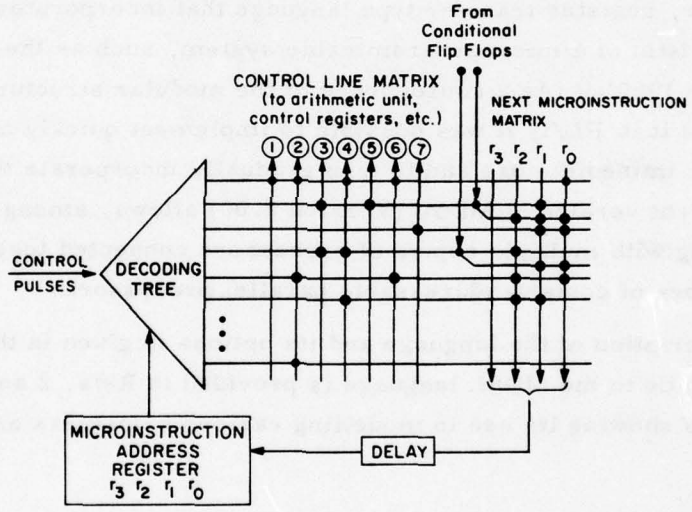


Fig. 2. Wilkes' model of microprogramming.

(1) Control Line Matrix. The output (vertical lines) of this matrix control the flow of data, instructions, and special control signals throughout the computer. For example, vertical line (1) might cause the program counter to be incremented. Each vertical line corresponds to a microoperation.

(2) Next Microinstruction Matrix. The output (vertical lines) of this matrix forms the binary address or label of the next microinstruction.

(3) Lines From Conditional Flip Flops. These inputs (vertical lines) contain information about the conditions and states of various parts of the computer.

(4) Decoding Tree. The inputs come from the Next Microinstruction Matrix and the output is a signal on one of the horizontal lines, the next microinstruction to be performed.

In Fig. 2 a dot at the intersection of a horizontal line and a control line means that the indicated microoperation will be performed as part of that microinstruction. The address of the next microinstruction is found by interpreting dots as 1's and no dots as 0's in the Next Microinstruction Matrix and then interpreting the 1's and 0's as a binary number. Note that the lines from the conditional flip flops cause some horizontal lines to be split into two lines before entering the Next Microinstruction Matrix. Thus, depending on the condition, one of two possible microinstructions will follow. Consequently, each microinstruction has a set of microoperations, a possible test, and one or two possible microinstructions to be performed next.

MDSL is organized as a set of relatively independent sections whose functions follow the Wilkes model. Each section begins with a control card which contains a $ in column one followed by one of the nine control section names.

$JOB, name

This, the first card in an MDSL job, causes a header page to be printed and all tables to be initialized.

$*DEF*

All definitions of information storage components: registers, flags, and memories follow this card.

$STATIC

This card precedes the definitions of the microoperations. This is where the controllable data flows are defined and given labels.

$MICRO

The "microprogram" of the computer-type device being defined is given following this control card. This corresponds to filling in the dots in the Control Line and Next Microinstruction matrices.

$INITIAL

After this the initial values of the registers and flags are given.

$PROGRAM

After this the initial values of various memory locations are given.

$RUN, TRACE(i, j, k, ···)(R1, R2, R3, ···), HIST

This causes the previously defined machine to be simulated, microinstruction by microinstruction with a trace that will cause the contents of registers R1, R2, R3, ··· to be printed out whenever microinstructions i, j, k, ··· are encountered. HIST causes a histogram showing the number of times each microinstruction was performed to be printed out upon successful termination of the simulation.

$DUMP

After this follows one or more memory dump specifications.

$END

This card is used to terminate the current job.

After a simulation has terminated, a user may simulate the running of a new program on the previously defined machine by simply following the $RUN card with a re-initialization of the key registers using $INITIAL, the entering of a new program using $PROGRAM, and a new simulation using $RUN. Because of its modular structure, MDSL allows control sections to be used and reused in a variety of ways.

C. Basics of the MDSL Language

The basic syntax and semantics of MDSL will be introduced by means of a set of constructs taken from the complete example given later in this report, where a set of processors asynchronously traverse a network in order to determine the minimum cost path. A detailed description of the syntax is given in Reference 3.

All information storage components (registers, flags, and memories) are defined in the $DEF section using the following basic format:

```
identifier : = storage specification
```

Below are three examples that illustrate three basic variations of the above general definition.

```
PTR : = <0: 8>    /POINTER TO THE UNASSIGNED MEMORY
```

In the above a 9-bit register with bit positions numbered from 0 to 8 is defined and assigned the name PTR. The / indicates the beginning of a comment; which will be assumed to run to the end of the current card.

```
MBR.8 : = <0: 7>    /MEMORY BUFFER REGISTER FOR BOTH MEMORIES
```

In the above a set of eight 8-bit registers with the generic name MBR are defined; each with bit positions numbered from 0 to 7.

```
PATH : = (0: 511)<0: 7>    /WILL CONTAIN THE PATHS AND PATH LENGTHS
```

In the above, a memory named PATH is defined to have 512 words (numbered from 0 to 511) of 8-bits each (numbered from 0 to 7).

In the $STATIC section all the controllable data flows are defined and given labels using the following basic format:

```
label:  source(s)> flow modifier> target(s)
```

This will cause the source(s) to be transferred to the target(s) under control of the flow modifier. The following examples demonstrate some of the possible variations of this basic form.

6: MBR>> NODE /SEND MBR TO NODE

The above defines control line 6 to control the transfer of the contents of register MBR into the NODE register.

7: 1>> SWT /RESET SWT TO 1

Control line 7 causes the constant 1 to be sent to the SWT register.

20: F.SWT> SET> F.SWT /SET FLAG OF CONTROLLED MACHINE

In the above, control line 20 controls the setting to one of the F registers of machine SWT.

13: MBR, COST> ADD> COST /UPDATE COST

If control line 13 is activated, then the contents of the MBR and COST registers will be added and the result will be stored in COST.

2: PATH(MAR)>> MBR /READ FROM PATH INTO MBR

Control line 2 controls the transfer of the contents of the location of memory PATH pointed to by the MAR register into the MBR register.

21: 00000000>> PATH(*) /CLEAR MEMORY INITIALLY

As defined above, control line 21 sets to zero all words of the memory named PATH. The * symbol is used to mean "all possible values."

The general format for a microinstruction defined in $MICRO is as follows:

label: (control line list)(condition)(branch specification)

The following examples demonstrate some of the variations of the above format.

2: (5, 3, 4) /READ CURRENT NODE, STORE IN PATH

The above microinstruction, labelled 2, causes control lines 5, 3 and 4 to be activated. Since no condition is specified, the next microinstruction will be the next one in sequence, that is, microinstruction 3.

11: ( )(F.SWT=1)(11) /LOOP IF CONTROLLED MACHINE IS NOT READY

Microinstruction 11 above will cause no control lines to be activated. It will check if the F register of machine SWT is 1. If it is, the next microinstruction will be number 11. Otherwise, the next will be number 12.

17: (3)(1)(0) /STORE COST AND HALT

Microinstruction 17 is a special case of a computed goto instruction. It causes

line 3 to be activated and then goes to the first branch address if the value in the condition field is 0, the second branch address if the value is 1, and etc.; with the last branch address being the goto address for all remaining values in the condition n field. In this case, there is an effective unconditional transfer to microinstruction 0. Transferring to microinstruction 0 halts the given machine. When all machines have halted, the simulation caused by the $RUN card terminates.

The $INITIAL and $PROGRAM sections enable one to initialize registers and memories. The following examples demonstrate the formats for these initializations.

```
$INITIAL
    C = 0   /INITIALIZE COST TO ZERO
```

The above causes the COST register to be initialized to 0. Unless otherwise stated, all constants are assumed to be binary.

```
    M.1 = 1   /MICRO-INSTRUCTION COUNTER OF MC.1 SET TO 1
```

This causes the microinstruction counter of machine 1 to be initialized to 1. The microinstruction counters of all machines are automatically built in and available for use in a microoperation or test condition. Because of this, it is possible to implement microsubroutines and complex interrupts in MDSL.

```
$PROGRAM
  supplied by
    MDSL      *TREE(1)
      1       00000001    /NODE NO.1
      ⋮          ⋮              ⋮
     25       00000000    /NO CHILDREN; TERMINAL NODE
```

The above will cause the memory TREE to be loaded with 0000001, ···, 00000000 starting with location 1 and going to location 25.

The general format for a dump specification in the $DUMP section is as follows:

name (lower address: upper address)

For example,

```
PATH(0: 99)   /DUMP OUT LOCATION 0 to 99
```

This will cause the contents of memory locations 0 through 99, inclusive, of memory PATH to be printed out. Unless OCTAL is specified on the $DUMP card, the output will be in binary.

D. Illustrative Example

An example which demonstrates some of the most significant features of MDSL is one in which an array of identical processors, operating asynchronously, examines the set of all paths through a loopless network in order to find the minimum cost path through the network. Figure 3 shows the network to be examined. A cost is associated with

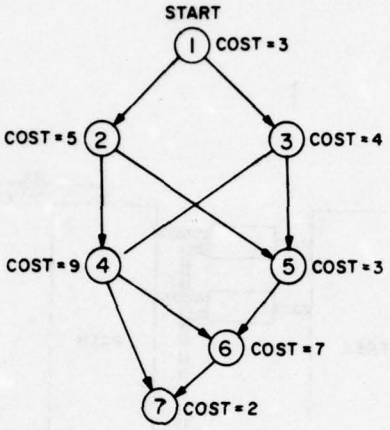


Fig. 3. Network to be followed.

each of the nodes. There is one start node, node 1 and one terminal node, node 7. An analysis of the network reveals that there are six unique paths from node 1 to node 7. The minimum cost path is found to be 1-3-4-7, with a cost of 18.

Figure 4 provides a diagram of the configuration of the machines and a detailed layout of a representative machine. (The MDSL simulation program is given at the end of this report.) The circled numbers associated with the various control lines in Fig. 4 correspond to the microoperation labels. For example, line 20 sets the F register of a machine to 1. This corresponds to the 20: F.SWT>SET>F.SWT entry in the $STATIC section of the MDSL program.

The multiprocessor example which required two simulations demonstrated the capability within MDSL of modifying and rerunning complex machines. The first simulation began with the description of the internal structure of one of the machines and the network to be followed. The definition allowed up to eight machines to be used. When the simulation was completed, it was found that only six machines were needed. Next the $DEF was used to add one register to the basic structure defined before, ZLIM; which would be used to contain the limit on the number of machines allowed to be used. In this way, we make the machine limit a variable. The use of ZLIM to limit the number

of machines was structured into the definition of the network by using the $MICRO to modify previously defined microinstruction six to read - 6: (8)(SWT = ZLIM)(8) /CHECK IF NO. OF MACHINES USED IS = LIMIT. After the reinitialization step, the modified network was simulated again. Since ZLIM was initialized to three, only three machines were allowed to follow the network. The results show that the same six unique paths were examined and the same minimum cost path was found. It should be noted that if ZLIM was initialized to two, an unresolvable deadlock would have developed in the network being simulated.

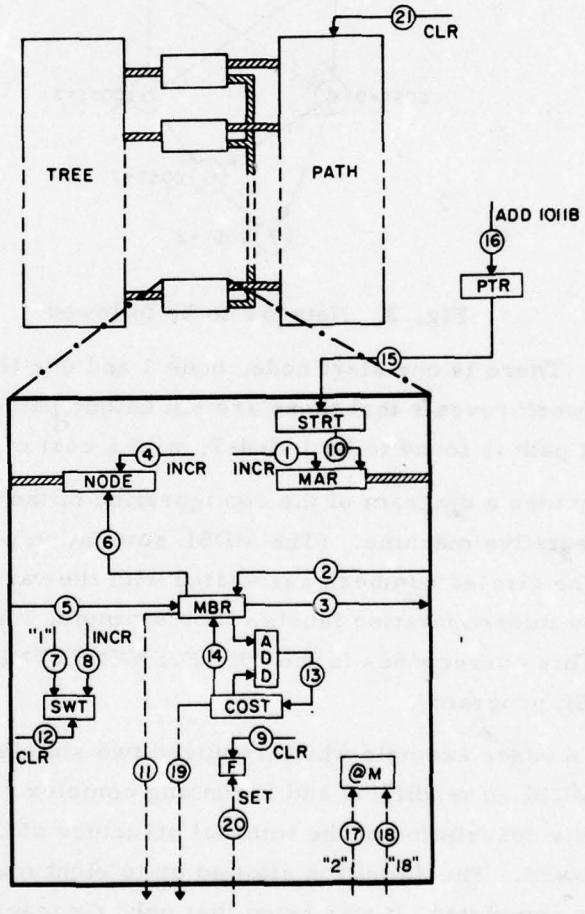


Fig. 4. Processor architecture.

E. Summary

The basic structure and syntax of the MDSL computer hardware design and simulation language has been described. The language is easy to learn and is based on the classic Wilkes model of a microprogrammable system. MDSL is currently being used in both an educational and a research environment. With MDSL a researcher can describe and simulate a variety of computers and computer-like structures, "debug" his design, and investigate the tradeoffs involved in various design alternatives. Two areas that are currently being explored involve the design and use of highly and loosely parallel structures, such as content addressable parallel processors and networks of processors, and the investigation of microprogramming structures.

MDSL is currently a developing language; new language constructs are still being added. Because of its modular form, it is relatively easy to add new features to the language in such a way as to allow for upward compatability.

Joint Services Technical Advisory Committee
F44620-74-C-0056

M. Adamowicz and J. Mirza

REFERENCES

1. M.V. Wilkes, "The Best Way to Design an Automatic Calculating Machine," Manchester University Inaugural Conf., pp. 16-18 (July 1951).
2. M. Adamowicz and J. Mirza, "Simulating Computer Hardware with MDSL," 1977 ACM Computer Science Conf., Atlanta, Georgia (February 1977).
3. M. Adamowicz and J. Mirza, "MDSL: A Microcomputer Design and Simulation Language," SIGMICRO Newsletter (June 1977).

MDSL Program for Example

```
$DEF

TREE := (0:127)<0:7>    / CONTAINS THE NETWORK TO BE FOLLOWED BY THE MACHINES
PATH := (0:511)<0:7>    / WILL CONTAIN THE PATHS AND THE PATH LENGTHS
NODE.8 := <0:8>         / HOLDS ADDRESS OF THE CURRENT NODE BEING VISITED
MAR.8 := <0:8>          / MEMORY ADDRESS REGISTER FOR THE MEMORY 'PATH'
MBR.8 := <0:7>          / MEMORY BUFFER REGISTER FOR BOTH MEMORIES
SWT.8 := <0:8>          / SWITCH THAT CHECKS FOR AN 'IDLE' MACHINE
COST.8 := <0:7>         / HOLDS THE COST ASSOCIATED WITH THE PATH
F.8 := <0>              / FLAG USED AS A SEMAPHORE
STRT.8 := <0:8>         / HOLDS STARTING ADDRESS OF THE ADDR. SPACE
PTR := <0:8>            / POINTER TO THE UNASSIGNED MEMORY


$STATIC

  1: MAR > INCR > MAR           / INCREMENT MAR
  2: PATH(MAR) > > MBR          / READ FROM PATH INTO MBR
  3: MAR > > PATH(MAR)          / WRITE INTO PATH FROM MBR
  4: NODE > INCR > NODE         / INCREMENT NODE
  5: TREE(NODE) > MBR           / READ FROM TREE INTO MBR
  6: MBR > > NODE               / SEND MBR TO NODE
  7: 1 > > SWT                  / RESET SWT TO 1
  8: SWT > INCR > SWT           / INCREMENT SWT
  9: F > CLR > F                / CLEAR FLAG F
 10: STRT > > MAR               / RESET MAR TO THE START OF IT'S ADDR. SPACE
 11: MBR > > ADR.SWT            / SEND MBR TO THE MBR OF ACTIVATED MACHINE
 12: SWT > CLR > SWT            / CLEAR SWT
 13: MAR, COST > ADD > COST     / UPDATE COST
 14: COST > > MBR               / SEND COST TO MBR TO BE STORED IN PATH
 15: PTR > > STRT               / DEFINE ADDRESS SPACE OF MACHINE
 16: (511,PTR > ADD > PTR       / ADD (511(ORIGINAL)) TO PTR
 17: 10 > > IM.SWT              / SET ACTIVATED MACHINE INTO INDEPENDENT MODE
 18: (0010) > > IM.SWT          / SET ACTIVATED MACHINE INTO CONTROLLED MODE
 19: COST > > COST.SWT          / INITIALIZE COST OF ACTIVATED MACHINE
 20: F.SWT > SET > F.SWT        / SET FLAG IF CONTROLLED MACHINE
 21: (000)0000 > > PATH(*)      / CLEAR MEMORY INITIALLY


$MICRO

  1: (21)                       / CLEAR MEMORY
  2: (5,3,4)                    / READ CURRENT NODE, STORE IN PATH
  3: (5,1)                      / READ COST ASSOCIATED WITH CURRENT NODE
  4: (13,4)                     / UPDATE COST
  5: (5,12)(9)<0> = 0((15)      / READ CHILD, CLEAR SWT, CHECK IF LAST CHILD
  6: (8)(SWT <= (0001)(8)       / INCREMENT SWT, CHECK IF WITHIN LIMITS
  7: (7)                        / RESET SWT TO ONE
```

```
 8: (3)(4M.SWT -= 0)(6)      / CHECK IF THE OTHER MACHINE IS ALREADY ON
 9: (10,11,13,18,19)         / CUT ACTIVATED MACHINE IN CONTROLLED MODE
                             / SET UP IT'S COST, ADDRESS SPACE, ETC.
10: (2)(MAR = 3)(13)         / READ PATH OF TO THE MOMENT
11: (3)(IF.SWT = 1)(11)      / LOOP IF CONTROLLED MACHINE IS NOT READY
12: (1,20,1)(1)(10)          / TRANSFER MAR, SET FLAG OF CONTROLLED MACHINE
13: (3)(IF.SWT = 1)(13)      / LOOP IF CONTROLLED MACHINE IS NOT READY
14: (4,17)(1)(5)             / SET ACTIVATED MACHINE INTO INDEPENDENT MODE
15: (3,6)(MODE -=0)(2)       / LAST CHILD IS FOLLOWED BY CURRENT MC.
16: (1,14)                   / PATH IS OVER; SEND COST TO MAR TO BE STORED
17: (3)(1)(0)                / STORE COST AND HALT
18: (15,6)                   / MACHINE IN CONTROLLED MODE
19: (10)                     / SET MAR TO START OF ADDRESS SPACE
20: (3)(IF=0)(20)            / LOOP TILL ACTIVATING MACHINE SENDS INFO.
21: (3,9)                    / STORE INFO; CLEAR FLAG
22: (1)(1)(20)               / LOOP BACK TO 20 FOR MORE INFORMATION
```

```
$INITIAL

 COST = 0                / INITIALISE COST TO ZERO
 NODE.1 = 4              / MACHINE ONE IS AT NODE ONE
 PC.1 = 1                / MICRO-INSTRUCTION COUNTER OF MC. 1 SET TO 1
```

```
$PROGRAM

          $TREE(1)
    1     00000001        / NODE # 1
    2     00000011        / COST = 3
    3     10000101        / ONE CHLD(N2) AT LOC. 5
    4     00011001        / LAST CHLD (N3) AT LOC 9
    5     00000010        / NODE 2
    6     00000101        / COST = 5
    7     10001101        / CHLD AT 13 (N4)
    8     00010001        / LAST CHLD (N5) AT 17
    9     00000011        / NODE 3
   10     00000100        / COST = 4
   11     10011101        / CHLD(N6) AT 13
   12     00010001        / LAST CHLD(N5) AT 17
   13     00000100        / NODE 4
   14     00001001        / COST = 9
   15     10010100        / ONE CHLD(N6) AT 20
   16     00010111        / LAST CHLD(N7) AT 23
   17     00000101        / NODE 5
   18     00000011        / COST = 3
   19     00010100        / ONLY ONE CHLD AT 20
   20     00000110        / NODE 6
   21     00000111        / COST = 7
   22     00011111        / CHLD (N7) AT 23
   23     00000111        / NODE 7
   24     00010010        / COST = 2
   25     00000000        / NO CHILDS, A TERMINAL NODE
```

```
$RUN, TRACE(1)(1)(NODE, COST, PTR), HIST

 AM.01 = 10     AM.02 = 18     AM.03 =  0     AM.04 =  0     AM.05 =  0     AM.06 =  0     AM.07 =  0     AM.08 =  0
   NODE.01 = 00000111  NODE.02 = 00000000  NODE.03 = 00000000  NODE.04 = 00001000  NODE.05 = 00000000  NODE.06 = 00000000
   NODE.07 = 00000000  NODE.08 = 00000000  COST.01 = 00000011  COST.02 = 00000011  COST.03 = 00000000  COST.04 = 00000000
   COST.05 = 00000000  COST.06 = 00000000  COST.07 = 00000000  COST.08 = 00000000  PTR = 000001011

 AM.01 =  9     AM.02 = 10     AM.03 = 18     AM.04 =  0     AM.05 =  0     AM.06 =  0     AM.07 =  0     AM.08 =  0
   NODE.01 = 00010111  NODE.02 = 00000111  NODE.03 = 00000000  NODE.04 = 00001000  NODE.05 = 00000000  NODE.06 = 00000000
   NODE.07 = 00000000  NODE.08 = 00000000  COST.01 = 00000111  COST.02 = 00001000  COST.03 = 00001000  COST.04 = 00000000
   COST.05 = 00000000  COST.06 = 00000000  COST.07 = 00000000  COST.08 = 00000000  PTR = 000010110

 AM.01 = 10     AM.02 = 11     AM.03 = 22     AM.04 = 18     AM.05 =  0     AM.06 =  0     AM.07 =  0     AM.08 =  0
   NODE.01 = 00010111  NODE.02 = 00000111  NODE.03 = 00001101  NODE.04 = 00000000  NODE.05 = 00000000  NODE.06 = 00000000
   NODE.07 = 00000000  NODE.08 = 00000000  COST.01 = 00000111  COST.02 = 00001000  COST.03 = 00001000  COST.04 = 00000111
   COST.05 = 00000000  COST.06 = 00000000  COST.07 = 00000000  COST.08 = 00000000  PTR = 000100001

 AM.01 = 15     AM.02 =  5     AM.03 = 10     AM.04 =  8     AM.05 = 18     AM.06 =  0     AM.07 =  0     AM.08 =  0
   NODE.01 = 00011110  NODE.02 = 00011001  NODE.03 = 00001111  NODE.04 = 00011111  NODE.05 = 00000000  NODE.06 = 00000000
   NODE.07 = 00000000  NODE.08 = 00000000  COST.01 = 00010001  COST.02 = 00010100  COST.03 = 00010001  COST.04 = 00010000
   COST.05 = 00010001  COST.06 = 00000000  COST.07 = 00000000  COST.08 = 00000000  PTR = 000101100

 AM.01 = 16     AM.02 =  0     AM.03 = 10     AM.04 = 10     AM.05 = 21     AM.06 = 18     AM.07 =  0     AM.08 =  0
   NODE.01 = 00010000  NODE.02 = 00000000  NODE.03 = 00001111  NODE.04 = 00001111  NODE.05 = 00010100  NODE.06 = 00000000
   NODE.07 = 00000000  NODE.08 = 00000000  COST.01 = 00010011  COST.02 = 00010100  COST.03 = 00010001  COST.04 = 00010000
   COST.05 = 00010001  COST.06 = 00010000  COST.07 = 00000000  COST.08 = 00000000  PTR = 000110111

                                    H I S T O G R A M

MACHINE  1

  1    1 |  2    5 |  3    5 |  4    5 |  5    7 |  6    6 |  7    0 |  8    6 |  9    2 | 10    5 |
 11    3 | 12    3 | 13    2 | 14    2 | 15    5 | 16    1 | 17    1 | 18    0 | 19    0 | 20    0 |
 21    0 | 22    0 |

MACHINE  2

  1    0 |  2    4 |  3    4 |  4    4 |  5    5 |  6    3 |  7    0 |  8    3 |  9    1 | 10    3 |
 11    2 | 12    2 | 13    1 | 14    1 | 15    4 | 16    1 | 17    1 | 18    1 | 19    1 | 20    2 |
 21    1 | 22    1 |

MACHINE  3

  1    0 |  2    2 |  3    2 |  4    2 |  5    3 |  6    5 |  7    0 |  8    5 |  9    1 | 10    4 |
 11    3 | 12    3 | 13    1 | 14    1 | 15    2 | 16    1 | 17    1 | 18    1 | 19    1 | 20    3 |
 21    2 | 22    2 |

MACHINE  4

  1    0 |  2    2 |  3    2 |  4    2 |  5    3 |  6    6 |  7    0 |  8    6 |  9    1 | 10    4 |
 11    3 | 12    3 | 13    1 | 14    1 | 15    2 | 16    1 | 17    1 | 18    1 | 19    1 | 20    1 |
```

```
 21    2 |  22    2 |

MACHINE  5

  1    0 |  2     2 |  3     2 |  4     2 |  5     2 |  6     0 |  7     0 |  8     0 |  9     0 |  10    0 |
 11    0 |  12    0 |  13    0 |  14    0 |  15    2 |  16    1 |  17    1 |  18    1 |  19    1 |  20    4 |
 21    3 |  22    3 |

MACHINE  6

  1    0 |  2     2 |  3     2 |  4     2 |  5     2 |  6     0 |  7     0 |  8     0 |  9     0 |  10    0 |
 11    0 |  12    0 |  13    0 |  14    0 |  15    2 |  16    1 |  17    1 |  18    1 |  19    1 |  20    4 |
 21    3 |  22    3 |

MACHINE  7

  1    0 |  2     0 |  3     0 |  4     0 |  5     0 |  6     0 |  7     0 |  8     0 |  9     0 |  10    0 |
 11    0 |  12    0 |  13    0 |  14    0 |  15    0 |  16    0 |  17    0 |  18    0 |  19    0 |  20    0 |
 21    0 |  22    0 |

MACHINE  8

  1    0 |  2     0 |  3     0 |  4     0 |  5     0 |  6     0 |  7     0 |  8     0 |  9     0 |  10    0 |
 11    0 |  12    0 |  13    0 |  14    0 |  15    0 |  16    0 |  17    0 |  18    0 |  19    0 |  20    0 |
 21    0 |  22    0 |

$DUMP

  PATH(0:99)          / DUMP OUT LOCATIONS 0 TO 99

  0 | 00000001 00000011 00000101 00000110 00000111 00000000 00010011 00000000 00000000 00000000 00000000
 11 | 00000001 00000010 00000101 00000110 00000111 00000000 00010100 00000000 00000000 00000000 00000000
 22 | 00000001 00000010 00000100 00000111 00000000 00010011 00000000 00000000 00000000 00000000 00000000
 33 | 00000001 00000011 00000100 00000111 00000000 00010010 00000000 00000000 00000000 00000000 00000000
 44 | 00000001 00000010 00000100 00000110 00000111 00000000 00011010 00000000 00000000 00000000 00000000
 55 | 00000001 00000011 00000100 00000110 00000111 00000000 00011001 00000000 00000000 00000000 00000000
 66 | 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000
 77 | 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000
 88 | 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000 00000000
 99 | 00000000

$OFF

ZLIM 3= <0:3>          / HOLDS LIMIT ON THE NUMBER OF MACHINES TO BE USED
```

never used

6 unique paths

Minimum cost path

cost

```
$MICRO

  G: IF(CNT <= ZLIM)[1]      / CHECK IF # OF MACHINES USED IS <= LIMIT


$INITIAL

ZLIM = 011
NCD(.1 = COST          / MACHINE ONE IS AT NCD, 1
COST = 0               / INITIALIZE COST TO ZERO
SM.1 = 1               / ACTIVATE MACHINE 1 AT MICRO-ORDER, 1


$RUN, HIST
```

HISTOGRAM

MACHINE 1

1	1	2	5	3	3	4	5	5	7	6	29	7	8	8	29	9	2	10	5		
11	3	12	2	13	2	14	2	15	5	16	1	17	1	18	0	19	0	20	0		
21	0	22	0																		

MACHINE 2

1	0	2	5	3	3	4	8	5	9	6	3	7	0	8	3	9	1	10	3		
11	2	12	2	13	1	14	1	15	8	16	3	17	3	18	3	19	3	20	10		
21	7	22	7																		

MACHINE 3

1	0	2	4	3	4	4	4	5	6	6	10	7	2	8	10	9	2	10	8		
11	6	12	0	13	2	14	2	15	4	16	2	17	2	18	2	19	2	20	6		
21	4	22	4																		

MACHINE 4

1	0	2	0	3	0	4	0	5	0	6	0	7	0	8	0	9	0	10	0		
11	0	12	0	13	0	14	0	15	0	16	0	17	0	18	0	19	0	20	0		
21	0	22	0																		

MACHINE 5

1	0	2	0	3	0	4	0	5	0	6	0	7	0	8	0	9	0	10	0		

Only 3 machines used

```
11    0 |  12    0 |  13    0 |  14    0 |  15    0 |  16    0 |  17    0 |  18    0 |  19    0 |  20    0 |
21    0 |  22    0 |
```

MACHINE 6

```
 1    0 |   2    0 |   3    0 |   4    0 |   5    0 |   6    0 |   7    0 |   8    0 |   9    0 |  10    0 |
11    0 |  12    0 |  13    0 |  14    0 |  15    0 |  16    0 |  17    0 |  18    0 |  19    0 |  20    0 |
21    0 |  22    0 |
```

MACHINE 7

```
 1    0 |   2    0 |   3    0 |   4    0 |   5    0 |   6    0 |   7    0 |   8    0 |   9    0 |  10    0 |
11    0 |  12    0 |  13    0 |  14    0 |  15    0 |  16    0 |  17    0 |  18    0 |  19    0 |  20    0 |
21    0 |  22    0 |
```

MACHINE 8

```
 1    0 |   2    0 |   3    0 |   4    0 |   5    0 |   6    0 |   7    0 |   8    0 |   9    0 |  10    0 |
11    0 |  12    0 |  13    0 |  14    0 |  15    0 |  16    0 |  17    0 |  18    0 |  19    0 |  20    0 |
21    0 |  22    0 |
```

$DUMP

PATH(0:99) / DUMP OUT LOCATIONS 0 TO 99

```
 0 | 00000001 C0000011 00000101 C0000110 00000111 C0000000 00010011 00000000 00000000 00000000 00000000
11 | 00000001 C0000010 00000101 C0000110 00000111 C0000000 00010100 00000000 00000000 00000000 00000000
22 | 00000001 C0000010 00000100 C0000111 00000000 C0000011 00000000 00000000 00000000 00000000 00000000
33 | 00000001 C0000010 00000100 C0000110 00000111 C0000000 00011010 00000000 00000000 00000000 00000000
44 | 00000001 C0000011 00000100 C0000111 00000000 C0000010 00000000 00000000 00000000 00000000 00000000
55 | 00000001 C0000011 00000100 C0000110 00000111 C0000000 00011001 00000000 00000000 00000000 00000000
66 | 00000000 C0000000 00000000 C0000000 C0000000 C0000000 00000000 C0000000 00000000 00000000 00000000
77 | 00000000 C0000000 00000000 C0000000 00000000 C0000000 00000000 00000000 00000000 00000000 00000000
88 | 00000000 C0000000 00000000 C0000000 00000000 C0000000 00000000 00000000 00000000 00000000 00000000
99 | 00000000
```

6 unique paths

Minimum cost path

cost

$END

CENTRALIZED TELEPROCESSING NETWORK DESIGN: A NEW HEURISTIC

R.R. Boorstyn and A. Kershenbaum

The problem considered is that of finding an optimal (minimum cost) design for a centralized telecommunication network given a set of locations, traffic magnitudes between these locations, and a single common source or destination.

Several heuristics, which are efficient (in terms of their execution time and memory requirements on a digital computer)[4,5,6,8,10] and which produce seemingly good results, have already been developed and are currently accepted techniques. Some work has also been done on finding optimal solutions[1,2,3] to this problem as a means of verifying the effectiveness of proposed heuristics. Kershenbaum and Boorstyn[7] have developed a technique which is capable of handling problems of realistic size by treating the problem within the context of matroid theory. On the basis of experiments run using this algorithm for the exact solution, an improved heuristic was developed which is shown in this report to yield significant improvements (over known techniques) in the quality of obtainable solutions while still requiring only a modest amount of computer time.

More formally, the problem is one of finding a minimum spanning tree subject to one or more constraints which in general are equivalent to demanding that the sum of the traffic associated with the nodes in any subtree must not exceed some predetermined maximum.

A minimum spanning tree is a loop-free collection of arcs joining a set of nodes such that the sum of the lengths of the arcs is minimal. In the case of a communication network, these collections of arcs are called multidrop lines.

This constraint form is quite general and encompasses many real-world constraints which arise in the design of centralized telecommunication networks. Thus for example, in addition to treating the obvious constraint imposed by line capacity, it is possible to treat a restriction on the number of terminals on a multidrop line by associating a uniform traffic with each terminal.

Also, the length (cost) functions which can be treated are quite general. Any function which is not a function of the tree chosen is permissible.

A. Problem Formulation

Given:

(1) A vertex (node) set $V = \{v_i \; i = 0, 1, \cdots, n\}$ representing the terminal locations in the network, where node v_0 is a distinguished node which we will refer to as the center,

(2) A symmetric distance function $D = \{d_{ij} \mid i, j = 0, 1, \cdots, N\}$ representing the cost (length of an arc), and

(3) A constraint, m, on the number of nodes which may share a multidrop line.

Minimize

$$C = \sum_i d_{ip_i}$$

subject to (3) where v_{p_i} is the immediate predecessor of v_i in T, i.e., the node closest to v_i on the path between V_i and V_0 in T and T is any spanning tree.

We define a multidrop line to be a subtree, T_j, rooted at the center, V_0, and define the set of nodes in T_j, to be V_j. Thus, the constraint can be stated as $|V_j| \leq m$ for every V_j, for the uniform traffic case and

$$\sum_{\substack{k \\ v_k \epsilon V_j}} TR_k \leq m$$

for the general case with TR_k the traffic associated with v_k.

We will refer to this problem as the capacitated minimum spanning tree (CMST) problem.

B. Heuristic Solutions

The most often used heuristic solutions to the CMST problem share the following properties: (1) their running time is a polynomial (usually of order between 2 and 3) in the number of nodes, N, and is independent of the constraint; (2) in the absence of constraints, they will yield a minimum spanning tree (MST); and (3) the quality of the solution (i.e., the amount by which it differs from the optimum) is not controllable and, except very loosely, is not known.

These heuristics can be placed in one of two classes; primal solutions, where one starts with a feasible solution to the problem and then enlarges or improves it; and dual solutions, where one starts with an (infeasible) MST and then generates modifications to make the solution feasible.

Primal algorithms are the more often used. Kershenbaum and Chou[8] have pointed out that all such algorithms fall into the class of MST problems constrained by traffic or response time requirements. The difference between them is mainly the sequential order with which a branch or a line is selected into the tree. Without the constraints, all the algorithms converge to a MST. With the constraints, they form different subtrees.

We describe below the Esau-Williams algorithm,[4] which is used in the new heuristic to generate solutions, and Karnaugh's Second Order Greedy Algorithm (SOGA),[6] which provided the motivation for the branch exchanges used in the new heutistic.

In the Esau-Williams algorithm, initially each node v_i is placed in a separate component, C_j. A tradeoff function, t_{ij}, is initialized for each arc, a_{ij}: $t_{ij} = d_{ij} - d_{i0}$; where d_{ij} is the distance or more generally the cost of an arc from v_i to v_j.

The arc, a_{i*j*}, for which t_{ij} is minimum is then sought and tested. If v_{i*} and v_{j*} are both in the same component or if the two components, when combined, violate the problem constraints, the arc is discarded and t_{i*j*} is set to a very large number. Otherwise, a_{i*j*} is added to the solution and the components containing v_{i*} and v_{j*} are merged. Values of t_{ij} are then recalculated by setting $t_{ij} = d_{ij} - d_{k0}$, where d_{k0} is the minimum, over all nodes ℓ in the newly formed component, of $d_{\ell 0}$. The procedure of seeking and testing arcs with minimum t_{ij} is repeated until a complete solution is found.

This procedure gives reference to nodes which are far from the center. Such nodes are difficult to treat in the sense that if one waits too long, it may be necessary to connect them directly to the center using a long arc.

In Karnaugh's proposed second order greedy algorithm, one iterates any primal CSMT algorithm, successively forcing selected arcs in or out of the solution, thus obtaining a sequence of solutions of improving quality. The increase in running time is a function of how many combinations of branches are forced in and out of solution. If one considers (potentially) all combinations, a branch and bound procedure results.

Karnaugh said little about how one selects the arcs to force in or out. The new heuristic presented here is essentially an outgrowth of this technique where a small subset of likely candidates for forcing are identified in the course of the algorithm. It thus becomes possible to obtain significant improvements in the quality of the solution without unreasonably increasing the algorithm's complexity beyond the point of practical applicability.

C. A New Heuristic

Recently, an improved technique for finding optimal solutions to CMST problems was developed.[7] By using the technique it was possible to obtain optimal solutions for a class of interesting, tightly constrained problems and make observations about characteristics of optimal solutions.

For a CMST problem with 18 nodes and $m = 3$, Fig. 1(a) is the solution yielded by the Esau-Williams algorithm. Figure 1(b) is the unconstrained MST on the same problem. Figure 1(c) is the optimal solution. Note that the two arcs present in the optimal solution, but not present in the Esau-Williams solution, are (G, O) and (B, K) and are MST arcs. While it is not necessarily so that all arcs present in the optimal solution and not in the Esau-Williams solution are MST arcs, in all the experiments run and checked to this end, invariably at least one arc in the optimum but not in the Esau-Williams solution was an MST arc.

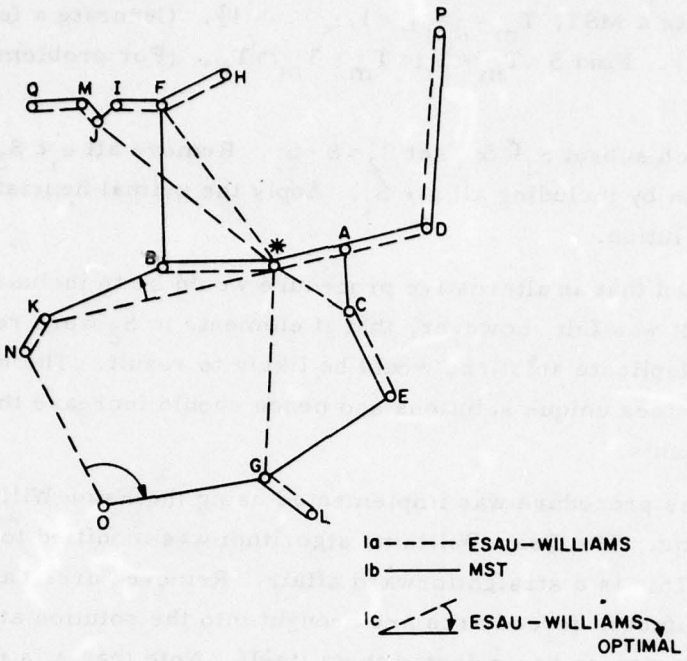


Fig. 1. Comparison between Esau-Williams' and optimal solution.

This is not to say that it is conjectured that one must always be able to find an optimum solution which contains an MST arc not part of the Esau-Williams solution (when the Esau-Williams solution itself is not optimal). It merely strengthens the point that MST arcs are reasonable arcs to examine when seeking to improve a heuristic solution.

Even without this evidence, however, it is intuitively appealing to consider such arcs. Most primal heuristics share the property that they proceed by connecting nodes to their nearest feasible neighbors.[8] The algorithms differ only in the order in which they consider the nodes. Thus, the solutions differ only in that the distance to a node's nearest feasible neighbor may differ depending upon when the node is considered during the execution of the algorithm. If such a heuristic fails to find the optimum, it will be because the algorithm erred by waiting too long before consideration of some critical node (or nodes) and hence missed being able to include some desirable arc (or arcs) in the solution. It is quite likely that one or more of these arcs will be MST arcs; indeed, as mentioned above, this was always the case with the experiments carried out.

Thus, a heuristic is proposed which attempts to improve solutions generated by some primal heuristic by forcing the inclusion of one or more MST arcs which the primal heuristic left out. This gives rise to the following procedure:

Step 1: Generate a MST, $T_m = \{a_j \mid j = 1, \cdots, n-1\}$. Generate a feasible tree, $T_f = \{b_j \mid j = 1, \cdots, n-1\}$. Find $S = T_m - T_f (= T_m - T_m \cap T_f)$. (For problems of interest, S is nonempty).

Step 2: For each subset $S_1 \subset S$: set $S_2 = S - S_1$. Remove all $a_j \in S_2$ from the network. Start a solution by including all $a_j \in S_1$. Apply the primal heuristic to complete the generation of a solution.

It should be noted that an alternative procedure would be to include a_j in S_1 but not to exclude a_j in S_2. It was felt, however, that if elements in S_2 were retained as candidates for inclusion, duplicate solutions would be likely to result. The exclusion of elements in S_2 guarantees unique solutions and hence should increase the likelihood of generating improvements.

A version of this procedure was implemented using the Esau-Williams algorithm as the primal heuristic. The Esau-Williams algorithm was modified to accept forced and removed arcs. This is a straightforward affair. Removed arcs have their lengths set to a very large number; forced arcs are brought into the solution at the start as if the Esau-Williams procedure has selected them itself. Note that it is possible for the forced arcs to imply infeasibility; this condition was tested for explicitly.

Problems were generated by reading in n and m and generating random x and y coordinates for the nodes within a unit square. The location of the center was random, centered, or in the corner, dependent upon the value of an input parameter. Euclidean distances were used.

A variety of problems were run with two goals:

(1) To see how the new heuristic compared with the optimum; and
(2) To see how the new heuristic compared with known heuristics.

Problems in the first class were relatively small ($n \leq 30$) and somewhat loosely constrained. In order to conserve computer time the problems were chosen from among those where the known heuristics performed most poorly. The new heuristic hit the optimum fairly consistently. In the thirty trials that were run, in only two cases did the new heuristic miss the optimal solution; in both cases, the new heuristic came within 1% of the optimum.

To see how the new heuristic compared with known heuristics, a broad range of problems for n ranging as high as 100 and m varying from n/3 to n/10 were examined.

Most primal heuristics generate subtrees which are MST's on the set of nodes they contain along with the center. Most of the arcs in these subtrees (with the exception of the arc directly connected to the center) will be MST arcs. Thus, one might expect

that as $|S| \approx \frac{n}{m}$, a vast improvement can be made over a blind branch exchange or SOGA procedure. However, this turns out to be slightly conservative. In practice, for $n \leq 20$ and v_0 in the geographic center, and for $n \leq 15$ and v_0 in the corner, $|S| \approx \frac{n}{m}$. For larger networks, $|S|$ ranges between $\frac{n}{m}$ and $\frac{n \log n}{m}$, where $|S|$ for networks with noncentered roots is generally larger than for the same network with v_0 in the center.

There are, in fact, $2^{|S|}$ subsets of S. Thus, for moderately large tightly constrained problems, S could grow large enough to make evaluating all subsets impractical. Furthermore, it is reasonable to assume that not all arcs in S interact with one another; i.e., improvements in separate parts of the network can be found and justified independently of one another. Further, in running the experiments, it was found that if a group of arcs yielded an improvement when forced into the solution, subsets of these arcs often yielded improvements too. Thus, the heuristic was modified by only considering subsets $S_1 \subseteq S$ such that $|S_1| \leq K$ for some given K. The best subset, S_1^*, is found and permanently forced into the solution. We then set $S = S - S_1^*$ and repeat the procedure until no further improvement can be made. As multiple branch exchanges are often required to effect improvements, the value of K chosen should be large enough to ensure that most advantageous exchanges will be found and at the same time small enough to keep the procedure computationally tractable; K = 2 seemed to work well. Experiments were run with larger values of K; only in isolated cases was any improvement over K = 2 obtained; in no case was an improvement in excess of 1% observed. Thus, it was decided in all the remaining experiments to use the extended procedure (which proceeds from the best solution found) and restrict the examination to subsets of cardinality less than or equal to two.

Using this modification of the new heuristic on a large number of problems, it was observed that in general, forcing arcs between nodes which are close to the center seemed to have the greatest effect on the value of the solution. This was probably a consequence of the fact that the Esau-Williams algorithm starts with nodes far from the center and hence, dealing with nodes near the center first radically changes the solution value.

Large subsets, in some cases all of S, sometimes yielded the best solutions. This means that one cannot simply consider subsets of limited size without iterating to subsequently force additional arcs into the solution.

In particular, it was found that for small networks ($n \leq 20$) with centered v_0, there was little difference (usually $\leq 1\%$) in performance between the new heuristic and the Esau-Williams solution. This is almost certainly because both procedures were generating near-optimal solutions. However, for larger networks (up to n = 100) and centered v_0, particularly for tightly constrained problems, the new heuristic performed

noticeably better with improvements averaging about 1.5%. Figure 2 shows the relative

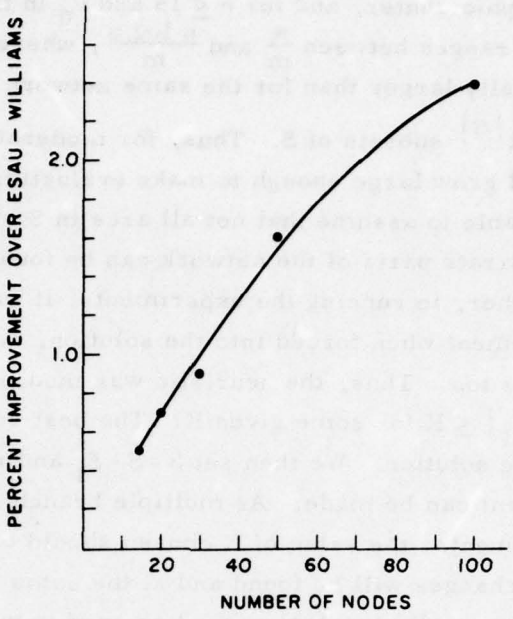


Fig. 2. Improvements obtained with the new heuristic.

improvement obtained using the new heuristic relative to the Esau-Williams algorithm. Each point on the curve represents an average over 12 sample runs. As can be seen, the improvement increases as the problem size does. Significant savings are obtainable for realistic problems.

Surprisingly, the improvement was only weakly correlated with problem tightness (ratio of m to n); the new heuristic achieved slightly larger savings in more tightly constrained problems. Related to this, is that substantial savings were achieved even for very loosely constrained problems. In such cases, the constrained solutions differed only slightly from the MST, as expected. What was startling was that the values obtained with the Esau-Williams algorithm sometimes differed from the MST value over 20% more than those obtained using the new heuristic; i.e.,

$$S = \frac{V_{ew} - V_{mst}}{V_{nh} - V_{mst}} = \frac{V'_{ew}}{V'_{nh}} \text{ was often over 1.2}$$

where

V_{ew} is the Esau-Williams value

V_{nh} is the new heuristic value

and

V_{mst} is the MST value.

Figure 3 is a plot of S vs. N, the number of nodes. Again, it is an increasing function of N.

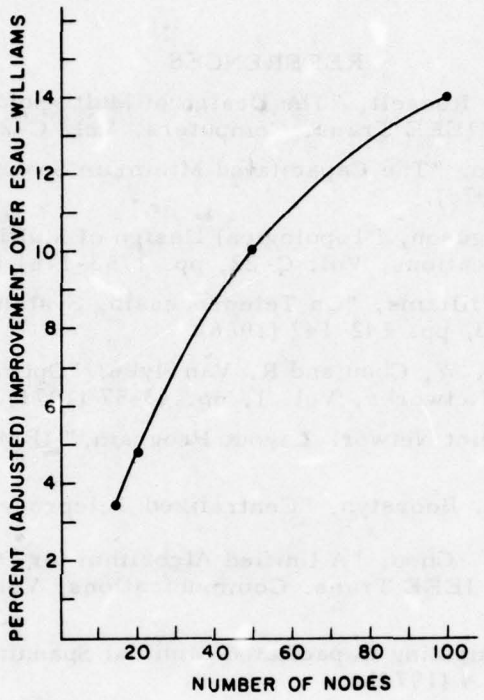


Fig. 3. Adjusted improvements using the new heuristic.

Results were even more encouraging for large networks with v_0 in the corner, such networks may be viewed as one-fourth of a network of 4n nodes with v_0 in the center. For such problems, improvements averaging 4% and as high as 8% were observed. Thus, the procedure appears to be of value for many realistically-sized problems with tight constraints.

A straightforward implementation of the Esau-Williams procedure has a computational complexity of order $n^2 \log_2 n$. A more careful implementation[9] can reduce the complexity to order $n \log_2 n$. As discussed previously, $|S|$ was found to range between $\frac{n}{m}$ and $\frac{n \log n}{m}$. Since we examine subsets of cardinality at most $\binom{|S|}{2}$ subsets are examined on each iteration. At worst, $\frac{|S|}{2}$ successive subsets of cardinality could be introduced and thus, the procedure could iterate at most $\frac{|S|}{2}$ times. In practice, the number of iterations grows more slowly than $|S|$.

Thus, a careful implementation of the procedure has a complexity of $|S|^3 n \log n$. If, for instance, $|S| \simeq n^{1/2}$, the procedure's complexity is $n^{5/2} \log n$.

Joint Services Technical Advisory Committee
F44620-74-C-0056 R.R. Boorstyn and A. Kershenbaum

National Science Foundation
ENG73-04232

REFERENCES

1. K.M. Chandy and R.A. Russell, "The Design of Multi-point Linkages in a Teleprocessing Tree Network," IEEE Trans. Computers, Vol. C-21, pp. 1062-1066 (1972).
2. K.M. Chandy and T. Lo, "The Capacitated Minimum Spanning Tree," Networks, Vol. 3, pp. 173-182 (1973).
3. D. Elias and M.J. Ferguson, "Topological Design of Multi-Teleprocessing Networks," IEEE Trans. Communications, Vol. C-22, pp. 1753-1761 (1974).
4. L.R. Esau and K.C. Williams, "On Teleprocessing System Design, Part 11," IBM Syst. J., Vol. 5, No. 3, pp. 142-147 (1966).
5. H. Frank, I.T. Frisch, W. Chou and R. Van Slyke, "Optimal Design of Centralized Computer Networks," Networks, Vol. 1, pp. 43-57 (1971).
6. M. Karnaugh, "Multipoint Network Layout Program," IBM, Report No. RC3723 (1972).
7. A. Kershenbaum and R. Boorstyn, "Centralized Teleprocessing Network Design."
8. A. Kershenbaum and W. Chou, "A Unified Algorithm for Designing Multidrop Teleprocessing Networks," IEEE Trans. Communications, Vol. C-22, No. 11, pp. 1762-1772 (1974).
9. A. Kershenbaum, "Computing Capacitated Minimal Spanning Trees Efficiently," Networks, Vol. 4, No. 4 (1974).
10. N.V. Reinfeld and W.R. Vogel, "Mathematical Programming," (Englewood Cliffs, N.J.: Prentice-Hall, 1958).

COMMUNICATION NETWORKS WITH ADAPTIVE CHANNEL CAPACITIES

K. Sohraby, L. Shaw and R. R. Boorstyn

We consider a network of nodes interconnected by communication channels. It is assumed that the route for each message from its origin node to its destination node has been chosen, and that at each node along this route the message waits in a buffer queue until the channel needed for its next step is free. The ith buffer has a maximum capacity of K_i messages so that messages arriving at a filled buffer will be blocked or lost, requiring retransmission at a later time.

The novel idea considered here is that the network's efficiency, in terms of average message delay and channel costs, can be reduced by allowing frequent changes in channel capacity. An expensive channel, with higher capacity might be employed when the buffer is nearly full; but a slower, less expensive channel might suffice when few messages await transmission. Such rapid changes, say after time intervals in which a small number of messages might arrive or depart, would be feasible in some channels currently being considered by communication engineers.[1] We refer here to a channel which is shared by voice and data users, with higher priority for the data. Adaptation of the capacity available for the data could be achieved at the expense of occasional reduction in the capacity available for the less important voice users.

Although one ultimate goal is the control of channel capacities for a network of interconnected queues, the first step will be to optimize the queue-dependent selection of channel capacities for a single queue.

The second step is the derivation of an effective short term departure rate for the approximate Poisson departure process from such a controlled queue. This departure model will then be used to describe the arrival process at the next queue in the network.

This report summarizes results achieved for these first two stages of the study, and then suggests the manner in which these results will be used for the overall network optimization.

A. Controlled Queue

A single buffer queue is modeled as an M/M/1/K queue i.e., with Poisson arrivals (rate λ), exponential service times, a single server, and a maximum storage of K messages. We use x_k to represent the state (number of messages waiting or in service) at time t_k. The capacity u_k used during the interval $(t_k, t_{k+1}]$ is to be a function of the state x_k at the beginning of the interval. Although the service

rate may change sequentially, the queue discipline is in all other respects just like the ordinary first-come, first-served one. This is in contrast to an alternate model[2] in which service in each interval is restricted to those messages already in the buffer at the beginning of the interval.

Three types of decision-time sequences are considered. The fixed interval control (FIC) procedure has equally spaced decision times $t_k = kT$. The random interval control (RIC) procedure assumes independent, identically distributed (i.i.d) inter-decision intervals, which might represent a situation where the decision computer is time-shared with other functions. Finally a continuous control (CC) procedure, which considers capacity changes after every arrival or departure, is analyzed for comparison (lower bound) with the other two more feasible cases.

Defining the sequence of decision intervals $T_k = (t_{k+1} - t_k)$, we have

FIC: $T_k = T$ all k

RIC: T_k i.i.d and independent of $\{x_j\}$

$$f_{T_k}(t) = \gamma e^{-\gamma t} \quad , \quad t \geq 0$$

CC: T_k i.i.d and indep of $\{x_j\}$

$$f_{T_k}(t) = \rho e^{-\rho t} \quad , \quad t \geq 0$$

where $\rho > \lambda + \mu_{max}$

In the CC case we have used the scheme of Lippman[3] to convert the original problem with state-dependent inter-decision intervals to one with state-independent intervals. In the converted model the mean interval durations are shorter, and there is a non-zero probability that the state does not change at each of these newly defined random times.

A cost criterion is essential if operational efficiency is to be optimized. The mean cost corresponding to operation over the interval T_k, starting in state $x_k = i$ is taken as

$$C_i(k) = \text{Mean message delay during } T_k + \alpha[\text{mean channel cost during } T_k] + \beta(\text{Expected number of blocked messages during } T_k) \quad (1)$$

Where α and β are weighting factors. Expressions for each of the terms in Eq. (1) will be defined separately for each of the three cases being studied. The ultimate

desire is to minimize the steady state mean loss rate, of the form

$$\bar{C} = \lim_{k \to \infty} \frac{1}{(t_k - t_o)} \; E\left\{ \sum_{r=0}^{k-1} C_{x_r}(r) \,\middle|\, x_0 = i \right\} \tag{2}$$

which will turn out to be independent of the initial state i.

Given a finite set of possible service rates $E = \{ \epsilon_1, \epsilon_2, \ldots \epsilon_M \}$, a stationary decision policy will be a function which assigns an ϵ_j to $\mu_i = \mu\,(x_k = i)$ for each $i = 0, 1, \ldots K$.

Since there is a finite number (M^K) of different policies, it is clear that an optimal one must exist to minimize $\bar{C}$. Howard's Dynamic Programming procedure[4] is an efficient way to find the best policy.

The analysis here is simplified by the fact that a stationary policy is assumed. Thus, the mean one-step cost $C_i(k)$ in Eq. (1) is not a function of k ; the sequence $\{ x_k \}$ is an ergodic Markov chain; and the loss rate in Eq. (2) is independent of the initial state $x_0 = i$. That limiting mean loss per step can also be written as

$$\bar{C} = \sum_{j=0}^{K} \pi_j C_j \tag{3}$$

with the aid of the steady state probabilities $\pi_j = P[\, x_k = j]$ defined by the conditions

$$\pi_j = \sum_{i=0}^{K} \pi_i P_{ij} \quad . \tag{4}$$

The stationary one-step transition probabilities P_{ij} will later be defined differently for each of the three cases.

Howard's procedure examines the mean total n-step loss

$$\begin{aligned} k_i(n) &= E\left[\sum_{k=0}^{n-1} C_{x_k}(k) \,\middle|\, x_0 = i \right] \\ &= C_i + \sum_{j=0}^{K} P_{ij}\, k_j(n-1) \; . \end{aligned} \tag{5}$$

In view of the limiting constant loss rate in Eq. (2), we have, for large n, approximately linear $k_i(n)$

$$k_i(n) \approx n\,\bar{C} + m_i \tag{6}$$

with intercepts m_i. Substitution of Eq. (6) into Eq. (5) leads to the set of (k+1) equations in (k+2) unknowns: $\overline{C}_1, m_0, m_1, \ldots m_k$.

$$\overline{C} = C_i + \sum_{j=0}^{K} P_{ij}\, m_j - m_i \; ; \; i = 0, 1 \ldots k \tag{7}$$

(C_i and P_{ij} will be computable for a given policy, and they are thus known when solving Eq. (7) for $\overline{C}$.) It is clear that Eq. (7) is unchanged if the same constant is added to every m_i, so we can set $m_0 = 0$ and reduce the number of unknowns to equal the number of linear equations.

Rather than solving Eq. (7) for $\overline{C}$ for each possible policy, Howard has shown that the followup iterative procedure conveys efficiently to the optimal policy.

Policy Improvement

For each i, find the $\epsilon_j = \mu_i$ which minimizes the test quantity

$$C_i(\mu_i) + \sum_{j=0}^{K} P_{ij}(\mu_i)\, m_j$$

in which the m_j from the previous policy are maintained. The resulting μ_i becomes a part of the improved policy, and the corresponding C_i and P_{ij} are used in the next value determination.

Value Determination

given C_i and P_{ij} for a given policy, solve (7) for $\overline{C}$ and m_i ($m_0 = 0$) .

The iteration is stopped, and the optimal policy achieved when no policy changes are found in the Improvement Routine.

B. Fixed Interval Control

In this case the instantaneous transition probabilities can be found as the solution of the birth-death equations for the finite storage queue.

$$\begin{aligned}
\dot{p}_o(t) &= -\lambda\, p_o(t) + \mu_i\, p_1(t) \\
\dot{p}_j(t) &= -(\lambda + \mu_i)\, p_j(t) + \lambda\, p_{j-1}(t) + \mu_i\, p_{j+1}(t); \; 0 < j < K \\
\dot{p}_K(t) &= -\mu_i\, p_K(t) + \lambda\, p_{K-1}(t)
\end{aligned} \tag{8}$$

where we identify $p_{ij}(\mu_i, t) = p_j(t)$ in Eq. (8) when the initial conditions are $p_i(0) = 1$ and other $p_j(0) = 0$, $j \neq i$. The one-step transition probabilities needed for Eq. (7) are thus $P_{ij} = p_{ij}(\mu_i, T)$.

In this FIC case, the one step mean cost from Eq. (1) becomes

$$C_i = \int_0^T \left[\sum_{j=1}^{K=1} j\, p_{ij}(\mu_i, t) + \beta\, p_{ik}(\mu_i, t) \right] dt + \alpha T\, C(\mu_i) \tag{9}$$

where $C(\mu_i)$ = cost/unit-time of a channel having service rate μ_i

μ_i = (mean bits/message) x (channel capacity).

C. Random Interval Control

Here the durations T_k are random so

$$P_{ij} = E[p_{ij}(\mu_i, T_k)] = \int_0^\infty \gamma e^{-\gamma t}\, p_{ij}(\mu_i, t)\, dt \tag{10}$$

using the $p_{ij}(\mu_i, T_k)$ generated by Eq. (8). Applying the expectation integral to both sides of Eq. (8) produces a set of simultaneous, linear, algebraic equations for the P_{ij}.

The C_i expression also simplifies greatly in this case because integration by parts yields

$$E\left[\int_0^{T_\ell} p_{ij}(\mu_i, t)\, dt \right] = \frac{1}{\gamma} P_{ij} \quad . \tag{11}$$

Thus

$$C_i = \frac{1}{\gamma} \left[\sum_{j=0}^{K-1} j\, P_{ij} + \beta\, P_{ik} + \alpha\, C(\mu_i) \right] \quad . \tag{12}$$

D. Continuous Control

Here we postulate Poisson event times with rate $\rho > (\lambda + \mu_{max})$ and transition probabilities

$$\begin{aligned}
P_{i(i+1)} &= \frac{\lambda}{\rho}\,; \quad 0 \le i < K \\
P_{i(i-1)} &= \frac{\mu_i}{\rho}\,; \quad 1 \le i \le K \\
P_{ii} &= 1 - (\lambda + \mu_i)/\rho\,; \quad 0 < i < K \\
P_{00} &= 1 - \frac{\lambda}{\rho} \\
P_{kk} &= 1 - \frac{\mu_i}{\rho} \\
P_{ij} &= 0 \text{ otherwise}
\end{aligned} \tag{13}$$

It can be shown that this process is equivalent to that of a channel-capacity-changing decision at every arrival and departure time. The stationary policies will clearly not change capacity at those event times where no state change occurs.

In this case, the expected one-step cost is

$$C_i = \frac{1}{p}\,(i + \alpha\, C(\mu_i)) \qquad i < K$$
$$= \frac{1}{\rho}\,(\beta + \alpha\, C(\mu_i)) \qquad i = K \tag{14}$$

E. Numerical Examples

Figure 1 compares the average cost per unit time $\overline{C}$, as a function of arrival rate λ, for several capacity-switching policies.

Here, the weighting parameter $\alpha = 0.3$, $\beta = 10$, the maximum number of messages which can be stored is $K = 4$, the available service rates are (0, 1, 2, 3, 4, 5, 6, 7, 8, 9), and the channel cost function is linear; $C(\mu) = \mu$.

Clearly, the lowest curve must correspond to the continuous control case, which is the most flexible. Moreover, the optimal policy in that case is of the bang-bang form: $\underline{\mu} = (0, 9, 9, 99)$. That is, we should use $\mu_0 = 0$ when $x_k = 0$ and the maximum $\mu_i = 9$ at all other times.

The middle curve in Fig. 1 corresponds to a fixed-interval controller with T = 1.

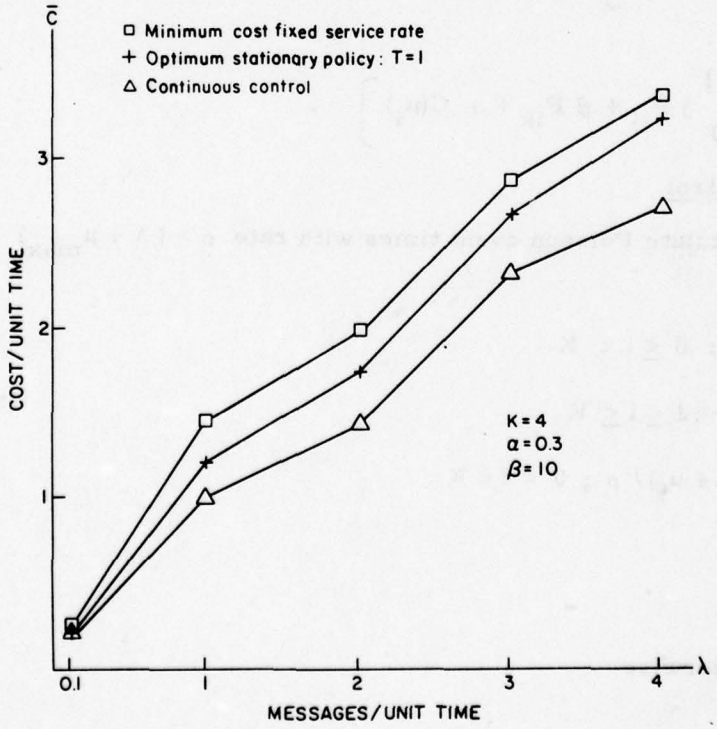


Fig. 1. Minimal cost for various capacity-switching disciplines.

The most costly curve in that figure results when capacity switching is not permitted, but the best fixed capacity is used for each arrival rate. Table I shows the optimal policies for these two cases, as functions of arrival rate.

TABLE I. Optimal service rates.

Arrival Rate λ	Optimal $\underline{\mu}$ Fixed - Interval	Optimal $\underline{\mu}$ No switching
0.1	(1, 2, 4, 5, 8)	(1, 1, 1, 1, 1)
1	(1, 3, 5, 7, 9)	(3, 3, 3, 3, 3)
2	(2, 4, 7, 7, 9)	(5, 5, 5, 5, 5)
3	(4, 6, 7, 9, 9)	(6, 6, 6, 6, 6)
4	(6, 7, 8, 9, 9)	(7, 7, 7, 7, 7)

These results were obtained via dynamic programming as outlined above. The fixed-interval case also required numerical integration of Eqs. (8) and (9) for various test policies.

F. Effective Departure Rate

When analyzing a network of interconnected, variable service rate queues, it will be necessary to characterize the departure process from each such queue during an optimization interval of duration T. As a fist step, we have defined and evaluated the mean departure rate η_i, for the fixed-interval case, as a function of the state i at the beginning of the interval.

If there were no limit on the size of the "waiting room" (K= ∞) we would define $\lambda \Delta t = P$ [an arrival in any Δt interval]

d_k = # of departures during interval T_k

to get

$$\eta_i = E \{ d_k \mid x_0 = i \} / T \tag{15}$$

$$= E \{ x_0 + \int_0^T \lambda \, dt - x_T \mid x_0 = i \} / T$$

or

$$\eta_i = \lambda - [E\{ x_T \mid x_0 = i \} - i]/T \tag{16}$$

It is well known that, for K = ∞, [5]

$$\mathcal{L}\{E\{x_T | x_0=i\}\} = \frac{s[i+\mu P^*_{i0}(s)] + \lambda - \mu}{s^2} \tag{17}$$

where

$$P^*_{i0}(s) = \mathcal{L}\{p_{i0}(t)\} .$$

Thus, for the $K = \infty$ case

$$\eta_i = \mu \left[1 - \frac{1}{T} \int_0^T p_{i0}(t)\, dt \right] \tag{18}$$

Curiously, Eq. (18) applies also for the finite - K case, as long as the appropriate $p_{i0}(t)$ is used (i.e., found from Equations (8)). To see this, we consider the modification of Eq. (15) to account for the probability

$$\lambda \Delta t \, [1 - p_{iK}(t)] = P[\text{an arrival in } (t, t+\Delta t) | x_o = i]$$

in place of $(\lambda \Delta t)$ for the unbounded case. In this way

$$\eta_i = \lambda - [\lambda \int_0^T p_{iK}(t)\, dt + E\{x_T | x_0 = i\} - i]/T \tag{19}$$

It has been found that the expectation in Eq. (19) has a Laplace Transform

$$\frac{s[i - \lambda P^*_K(s) + \mu P^*_0(s)] + \lambda - \mu}{s^2}$$

It follows that Eq. (18) is valid here also.

The general shape of η_i vs. i is monotone increasing toward an asymptote of $\eta_i = \mu$. If "many" messages are present at the outset $(x_0 > \mu T)$, then messages will be departing continuously at the channel capacity rate of μ .

G. Network Design

In a network, the arrival process at one queue will be formed as a combination of departure processes from other queues. The intention is to optimize the network policy by sequentially optimizing the individual queues, with updating of the arrival parameters at each step based on charges in the departure processes from the other queues.[6]

Joint Services Technical Advisory Committee
F44620-74-C-0056

K. Sohraby, L. Shaw and R.R. Boorstyn

REFERENCES

1. N. Shimasaki, T. Kohashi, T. Harbara and Y. Suzuki, "A Compatible Multiplexing Technique for Anisochronous and Isochronous Digital Data Traffic," Data Networks, Analysis and Design; Third Data Communications Symposium, St. Petersburg, Florida, pp. 59-67 (November 1973).

2. W.C. Chan, "Computer-Controlled Queueing Systems with Constant-Excess Cycle and General Service Times" IEE Proc. 117, pp. 927-930 (1970).

3. S.A. Lippman, "Applying a New Device in the Optimization of Exponential Queueing Systems," Operations Research, 23, Vol. 4, pp. 687-710 (July 1975).

4. R.A. Howard, "Dynamic Programming and Markov Processes," Technology Press, M.I.T. (1960).

5. J.W. Cohen, "The Single Server Queue," Wiley (1969).

6. C.B. Brosilow, L.S. Lasdon and J. D. Pearson, "Feasible Optimization Methods for Interconnected Systems" Proc. of Joint Automatic Control Conf. pp. 79-84 (1965).

UPPER BOUNDS ON THE NUMBER OF TESTS NEEDED TO VERIFY A COMPUTER PROGRAM

G.S. Popkin

The upper and lower bounds on the minimum number of program test cases were discussed in Ref. 1, with application to flowcharts containing two-way decisions. The conditions for reaching the lower and upper bounds were discussed, and examples given. In this report, these ideas are extended to flowcharts containing three-way (e.g., A < B, A = B, A > B) and multi-way decisions.

A. Upper and Lower Bounds on the Number of Tests Needed to Verify a Program

Each of the two flowcharts in Fig. 1 contains four three-way decisions. The numbering of the segments, with two segment numbers on some of the flow lines, indicates that the decisions are three-way. In Fig. 1(a), the methods of Reg. 2 would yield a maximum incomparable set size (and hence a lower bound on the minimum number of tests) of 3. It will be shown below how the upper bound may be computed, and how the flowchart contents can be inserted to raise the minimum number of tests required to approach the upper bound.

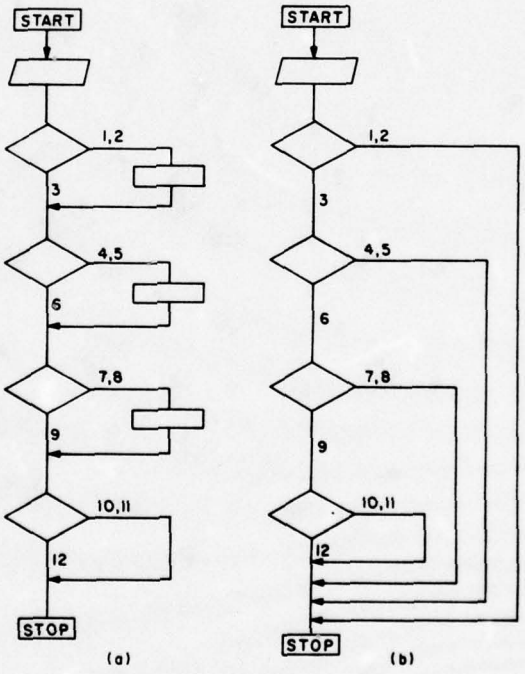


Fig. 1. (a) A flowchart with the maximum incomparable set of size 3.
(b) A flowchart with the maximum incomparable set of size 9.

In Fig. 1(b), the methods of Ref. 2 yield 9 as the size of the maximum incomparable set, and the lower bound on the minimum number of tests. 9 is also the upper

bound, for no flowchart contents can raise the minimum number of required tests above 9.

1. Minimum Number of Tests for Charts with Three-Way Deciders

In a loopless flowchart with three-way decisions, the upper bound on the minimum number of tests needed to pass through each segment at least once is given by

$$u = 2d + 1$$

where d is the number of deciders in the flowchart.

Proof: Consider a flowchart with no deciders. Such a flowchart consists of one segment and requires one test. Each three-way decider added to the flowchart can require at most two additional tests, so $u = 2d + 1$.

2. Minimum Number of Tests for Charts with Multi-Deciders

In a flowchart where the deciders may have any different numbers of outcomes, the upper bound on the minimum number of tests needed to pass through each segment at least once is given by

$$u = \sum_{i=1}^{n} w_i - n + 1$$

where w_i is the number of outcomes of decider i, and n is the number of deciders in the flowchart.

Proof: Consider a flowchart with no deciders. It consists of one segment and requires one test. Each decider i can require at most $w_i - 1$ additional tests, so

$$u = \sum_{i=1}^{n} (w_i - 1) + 1$$

$$= \sum_{i=1}^{n} w_i - n + 1$$

as asserted.

Figure 2 portrays the flowchart of Fig. 1(a) with contents. The minimum number of tests needed to pass through each segment at least once is now 7. The computed upper bound for the flowchart is 9.

In the deciders in Fig. 2, the segment numbers have the following meanings:

Segment No.	Outcome
1	P = 8
2	P > 8
4	P = 5
5	P > 5
7	P = 2
8	P > 2
10	P = 0
11	P < 0

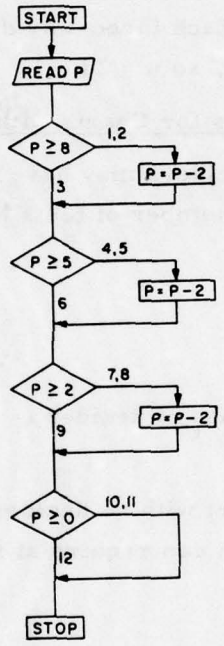


Fig. 2. A flowchart with contents.

If the input variable P may take on the seven values 1, 2, 4, 5, 7, 8, and 9, then the flowchart of Fig. 2 would require seven tests to traverse each segment at least once.

P	Path Traversed
1	3-6-9-11
2	3-6-7-10
4	3-6-8-12
5	3-4-8-12
7	3-5-8-12
8	1-5-8-12
9	2-5-8-12

The above illustrates the calculation of the upper bounds on the number of tests. Work is continuing on obtaining the actual number of tests rather than just a pessimistic upper bound.

Rome Air Development Center
F30602-74-C-0294

G.S. Popkin

REFERENCES

1. G.S. Popkin, "Program Paths and the Minimum Number of Tests Needed to Verify a Computer Program," Report No. POLY-EE/EP-76-003, Software Modeling Studies, July 1-December 31, 1975, Polytech. Inst. of New York.

2. M. Lipow, "Application of Algebraic Methods to Computer Program Analysis," Report No. TRW-55-73-10, TRW Software Series (May 1973).

EXTENSIONS OF SOFTWARE PHYSICS TO MEASURES OF COMPLEXITY

E. Berlinger and H. Ruston

In his work on software physics, M. Halstead[1] introduced a measure of complexity based upon certain program parameters. With the parameters:

N_1 = number of distinct operators

N_2 = number of distinct operands

the measure is

$$V = N_1 \log_2 N_1 + N_2 \log N_2$$

which he calls the program volume. Additional definitions and theorems are also given or conjectured. These relate the effort in programming, the total programming time, and estimate the program length.

Mathematically, the formulas are empirical and the proofs heuristic. But the results verify well with experimental data. The present effort attempts to improve on the Halstead scheme.

A. Outline of Current Work

One of the principal purposes of the work is to refine the definition of program volume to make it mathematically more sound, and also to include frequency of usage of the various program constructs. If we define:

f_1 = frequency of usage of the ith operator

p_i = probability of usage of the ith operator

f_j = frequency of usage of the variable whose rank is j

p_j = probability of usage of the variable whose rank is j

we can then define a measure of complexity as

$$- \sum f_i \log_2 p_i - \sum f_j \log_2 p_j \quad .$$

There is strong justification for this definition from an information theory point of view. This measure should correlate well with the number of bugs in a program. If so, then the number of bugs can be predicted from an initial version of the program.

Work is currently focused on automating the process for gathering the statistics necessary to obtain the p_i and p_j. To this end, the operating system of an IBM 370/125 is being modified to copy all error-free FORTRAN student programs onto a tape. Student programs collected over a full semester will then yield the necessary probabilities.

Statistics on errors will also be collected and automatically copied onto a second tape. Specifically, the FORTRAN error numbers will be obtained from the output queue before printing. These will give the syntax and run-time errors. To obtain a count of logical errors, a count of the total number of times a program is run is being kept. It is assumed that the number of logical errors is one less than the number of runs which yield no syntax or run-time errors. This may be an underestimate but since the programs being used are of first year programming students, it is reasonable to assume that they only find one logical error at a time.

To obtain the frequency counts, the tape containing the error-free runs will be run through a program supplied by Professor Halstead and Mr. Ottenstein of Purdue, which analyzes FORTRAN programs and yields the frequencies automatically. All programs on both tapes will be sufficiently identified so that the bugs can be correlated with the frequency counts.

Obtaining the probabilities p_i and p_j is a secondary purpose of the project and will supplement some statistics obtained previously by D. Knuth.

B. Tests for Obeyance of Zipf's Law

It is expected that the probabilities p_i and p_j will also obey Zipf's law, either in its pure form (i.e., $p_r = c/r$, where p_r is the probability of the operator or operand whose rank is r), or in one of its modified forms (e.g., $p_r = c/(r + a)^n$). If the frequencies also follow a Zipf's law, it may be possible to get a criterion for program length. This, however, remains to be seen.

Rome Air Development Center
F30602-74-C-0294

E. Berlinger and H. Ruston

REFERENCES

1. M. Halstead, "Software Physics, Basic Principles," IBM Research Report, RJ1582, IBM Research, Yorktown Heights, N.Y. (May 1975).

COMPLEXITIES OF NATURAL AND COMPUTER LANGUAGES

M. L. Shooman and A. E. Laemmel

There is a great need for theoretical models which describe programs and allow us to quantitatively estimate complexity, running time, storage requirements, and development time. In addition to serving as an estimate during the initial design period, they can be refined as the program develops and used as a management and analysis tool. They can also be used to compare initial design approaches, programming styles, different algorithms, etc. Early work on such a theory has been initiated.[1] This work discusses the linguistic theory (Zipf's laws), extends these to programming languages, and develops equations for program length based on these principles. This work is similar to that of Halstead which is commonly known as "Software Physics."[2]

There are many similarities between natural and computer languages, and we will make use of the analogies between nouns and verbs and operands and operators. The similarities in structure and content of natural and computer language are further illustrated via the following thought problem. Suppose we take a programmer who understands a particular computer language and give him the complete listing, comments, and documentation for a computer program. We instruct him to study the computer program until he understands it and then produce a report written in good English containing paragraphs, complete sentences, and algorithms written as a sequence of steps in good English without mathematical notation. In principle, the report and the computer program would be equivalent.

A. Zipf's Law

Before we discuss Zipf's law it is convenient to introduce a few terms in dealing with natural language. We use the term token to refer to all the words of the written or spoken sample. The term type is used to refer to the vocabulary of words in the sample. Much of our efforts will be centered on the counting of the number of times, n_r, particular types occur in a sample of n tokens containing t types. The most frequently occurring type will be assigned rank r = 1, the second most frequent type rank r = 2, and the least frequent type rank r = t. Thus

$$\sum_{r=1}^{t} n_r = n \quad . \tag{1}$$

The absolute frequency of occurrence for type r is n_r; however, the relative frequency of occurrence f_r is simply n_r/n.

Zipf studied the relationship between relative frequency of occurrence f_r and rank r for words from English, Chinese, and the Latin of Platus.[3]

Careful study of Zipf's data and that of others shows that f_r vs. r plots as a straight line on log-log paper, with a unity slope, thus we arrive at the simple relationship called Zipf's law,

$$f_r \cdot r = c \tag{2a}$$

$$n_r = \frac{cn}{r} \tag{2b}$$

Inspection of Eq. (2a) yields the fact that the constant c can be interpreted as the relative frequency of the rank 1 word type (also the intercept with the r = 1 line).

B. Type Token Equation

If we sum both sides of Eq. (2b), we obtain, using Eq. (1),

$$n = \sum_{r=1}^{t} n_r = cn \sum_{r=1}^{t} \frac{1}{r} \quad . \tag{3}$$

The summation of the series 1/r is given by[4]

$$\sum_{r=1}^{t} \frac{1}{r} = 0.5772 + \ell n\ t + \frac{1}{2t} - \frac{1}{12t(t+1)} \cdots \tag{4}$$

Substitution from Eq. (4) into Eq. (3) (retaining only 2 terms for modest size t) yields

$$n \approx cn(0.5772 + \ell n\ t) \tag{5}$$

We can eliminate the constant c from Eq. (5) by considering the behavior of Eq. (2b) for the smallest rank which is where $r_{max} = t$ (e.g., if there are 100 types, then the largest rank is obviously 100). In most cases the rarest type (largest rank) will occur only once, thus, $n_{r_{max}} = 1$. Substituting these values yields, $c = t/n$, which when combined with Eq. (5) gives

$$n = t(0.5772 + \ell n\ t) \tag{6}$$

C. Summary of Experimental Results

The results to date have shown that both operators, operands, and the sum of operators plus operands rectify fairly well with a slope of unity on log-log paper (i.e., they fit Zipf's law) for:

(1) An 11 line and a 27 line PL/1 program (55 and 222 tokens)

(2) The MIKBUG machine language executive program for the M6800 microprocessor (322 tokens)

(3) Operators in PDP-11 assembly language programs (1572 tokens)

(4) Variable names in 3 PL/1 programs (320, 238, and 193 tokens).

D. Relationship to "Software Physics"

The initial motivation for the application of Zipf's law to computer languages came from a review of Halstead's work[2] on Software Physics. Early in his work he arrives at a formula for program length

$$L = N_1 \log_2 N_1 + N_2 \log_2 N_2 \tag{7}$$

where

$L \equiv$ Program length

$N_1 \equiv$ Number of distinct operator types

$N_2 \equiv$ Number of distinct operand types

In terms of our notation the analogous quantities are

$$t = N_1 + N_2 \tag{8}$$

$$n = L$$

Note that Eqs. (7) and (6) are of similar form. In Ref. 1 we compare the actual number of tokens (counted) with the number of tokens calculated using both Equations (6) and (7). The average error and average magnitude error are computed, and both equations yield good agreement (10-20%), between actual and calculated results.

E. Estimation of Program Length Early in Design

One method of initially estimating program length (token length*) is to estimate the number of tokens. We assume the analyst initially has a complete description of the problem and that a partial analysis and choice of key algorithms has been made. An elementary approach might be to estimate the token size by

(1) Estimating the number of operator types which will be used in the language by the assigned programmers.

(2) Estimate the number of input variables, output variables, intermediate variables, and constant need.

(3) Sum the estimates of steps (1) and (2) and substitute in Equation (6).

Clearly we might consult past programs written by the assigned programmers for data on number of operator types. Also, if a large program will be stated in a comprehensive specification document, written in English or in a specific alien language. This

*In addition, one must add other classes of statements and programming elements such as: comments, declares, certain assembler directives, etc.

should name an input and output variable, thus our estimates will mainly deal with predicting the number of intermediate variables and constants.

Rome Air Development Center
F30602-74-C-0294

M.L. Shooman and A.E. Laemmel

REFERENCES

1. A. Laemmel and M. Shooman, "Statistical (Natural) Language Theory and Computer Programs," Report No. POLY-EE/EP-76-020, Polytechn. Inst. of New York (January 1977).

2. M. Halstead, "Software Physics, Basic Principles," IBM Research Report, RJ1582, IBM Research, Yorktown Heights, N.Y. (May 1975).

3. G.K. Zipf, "The Psycho-Biology of Language: An Introduction to Dynamic Philology," M.I.T. Press (1965). (First Houghton Miffin Edition, 1935.)

4. L. Jolley, "Summation of Series," (New York: Dover Publications, 1961), p. 36, n. 200 and p. 14, n. 70. (Note the constant 0.5772 is called Eulers constant, see "Differential and Integral Calculus," R. Courant (New York: Interscience Publishers, 1951) Vol. 1, pp. 381, 420.)

EXPERIMENTAL VERIFICATION OF DEBUGGING MODELS

D. L. Baggi

The object of this study is the implementation of a so-called driver program which, given a program to test, will force the traversal through all its possible paths. The advantage of such a procedure is obvious for programs with several branches and decision points; they are usually debugged by laborious construction of a data set, which hopefully would cause exploration of all paths. The method described here, forces exhaustive testing of all possible paths, with no need for the design of a data set.

A. Initial Drivers

The first effort in the definition of the driver program consisted of the implementation of a program capable indeed of traversing all paths of a given program. This program iteratively substitutes, in place of the condition in a PL/1 IF-statement, a value of zero or one, alternatively. With this done concurrently for all conditional statements, eventually one traverses once all possible paths. It was realized[1] that the resulting number of paths of 2^n for conditional branches, is merely an upper bound for all possible paths, where the real number lies in fact, between $n + 1$ and 2^n. Hence the above scheme wastes execution time, which increases exponentially (for example, a program with fourteen paths may require 8142 runs, of which 8128 are meaningless!). Thus an algorithm had to be designed to consider the possible paths only.

B. Present Drivers

The present algorithm scanns a PL/1 program. It searches for keywords such as IF, THEN, ELSE, DO and END, while constructing regular expressions of zeros and ones. Resolution of this regular expression yields a set of binary integers, which represent the status of the conditional expressions during each of the runs through all possible paths. In fact, each bit of such integers represents the value of the next conditional expressions met during execution, hence forcing a zero or one-branch accordingly. Since those integers were derived from the very structure of the algorithm of the program, they represent indeed each path; as an extra bonus, their total number is the number of possible paths. Hence the algorithm enumerates each path, uniquely describing it in terms of its branches, and also counts them.

The algorithm works as follows:
each expression is binary; it contains two terms separated by a + sign
the terms can be only 1,0, or 1 or 0 concatenated with a binary expression.

Scanning rules:

(a) each IF opens a left parenthesis, (

(b) each THEN corresponds to a 1

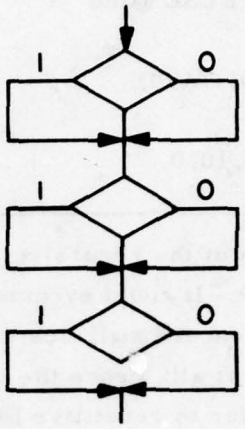
(c) each ELSE corresponds to a 0

(d) each well completed binary expression, with both terms completed, compels closing with right parentheses at its level.

Note: The algorithm takes care of missing ELSE clauses.

C. Examples

```
IF cond THEN stmt;
       ELSE stmt;
IF cond THEN stmt;
       ELSE stmt;
IF cond THEN stmt;
       ELSE stmt;
```



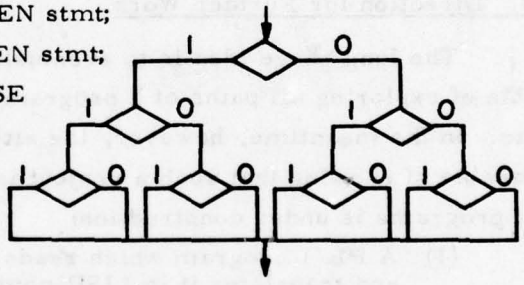
Result:

(1+0)(1+0)(1+0)

or the eight paths

111, 011, 101, 001, 220, 010, 100, 000

```
IF cond THEN IF cond THEN IF cond THEN stmt;
                                  ELSE stmt;
                        ELSE IF cond THEN stmt;
       ELSE IF cond THEN IF cond THEN stmt;
                        ELSE IF cond THEN stmt;
                                  ELSE
```



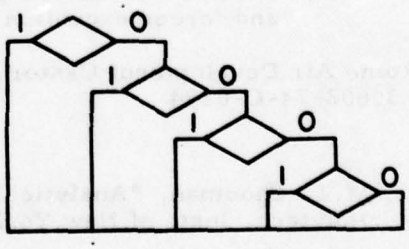
Result:

(1(1(1+0)+0(1+0))+0(1(1+0)+0(1+0)))

which gives the eight paths

111, 110, 101, 100, 011, 010, 001, 000

```
IF cond THEN stmt;
       ELSE IF cond THEN stmt;
                        ELSE IF cond THEN stmt;
                                  ELSE stmt;
```



Yields:

(1+0(1+0(1+0(1+0))))

or

1, 01, 001, 0001, 0000

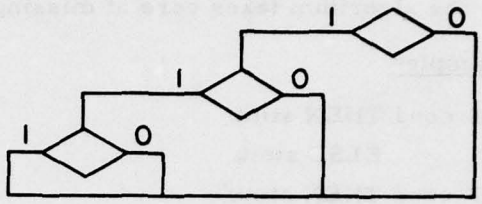
IF cond THEN DO; IF cond THEN DO; IF cond THEN stmt;
ELSE stmt;
END;
ELSE stmt;
END;
ELSE stmt;

Gives:

(1(1(1+0)+0)+0)

i.e.,

111, 110, 10, 0



Because of the recursive nature of the algorithm, it has been implemented in the language LISP. It could eventually be translated in PL/1. In the meantime, however, the greatest concern still lies in making sure that such an approach, and such an algorithm, works at all; hence the choice of LISP for fast implementation. No attention was given so far to repetitive DO-groups, which were appropriately handled by the initial crude algorithm. This is because extension of this algorithm to such loops is conceptually very trivial, and it can be proved that the system would perform equally well; but, in the meantime, such extension looks very time-consuming and would by no means add any contribution to the theory of this debugging model.

D. Direction for Further Work

The long range idea is to eventually come up with a complete PL/1 package, capable of exploring all paths of a program, paying attention to IF-statements, DO-loops, etc. In the meantime, however, the attention is given to producing a tentative system capable of showing that such a project is indeed possible. To this end, the following set of programs is under construction:

(1) A PL/1 program which reads in the object program (the one under debugging) and translates it in LISP-compatible notation; this is saved on a file.

(2) A LISP program which scans the object program and constructs the regular expressions, with results saved on a second file.

(3) A PL/1 with driver program which reads the results of these expressions and forces execution of the object program through all its possible paths.

Rome Air Development Center
F30602-74-C-0294

D.L. Baggi

REFERENCES

1. M.L. Shooman, "Analytical Models for Software Testing," Unpublished Research, Polytech. Inst. of New York.

AUTOMATIC PROGRAMMING TECHNIQUES

E. Lipschitz

Different methods and techniques for writing a better software package are currently being sought. The goal is to write programs which require a shorter testing and debugging time to achieve a certain degree of reliability.

One approach is automatic programming. The use of pre-written and already tested code modules reduces the number of bugs.

A. The Program "AUTO-PROGRAMMING"

The working hypothesis is that there exists a high degree of commonality among commercial applications which can be exploited to automate the production of code, once processing and output specifications are defined.

"AUTO-PROGRAMMING" is divided into two parts -- "Flow" and "Auto," both of which are interactive on-line programs that, by communicating with the user, generate his programs.

1. The Program "Flow"

"Flow" receives the information about the flow-chart of a program from the user and generates it. "Flow" recognizes only four different types of blocks, which are sufficient to generate any flow-chart. They are:

Type No. 1: Control block: A conditional decision block, similar to the statement If (_____). Go to (_____).

Type No. 2: Functional block: This block will perform a complete task selected from those in the computer library.

Type No. 3: Stop block: This block indicates the end of a path; i.e., Stop statement.

Type No. 4: User's code block: The user inserts the code he wants into this block. This feature is used whenever the libraru does not include programs for the needed task.

Upon completing the flow-chart, control passes to "Auto." "Auto" will generate the code for blocks Types No. 2 and No. 3, while the user will generate the code for blocks Types No. 1 and No. 4.

2. The Program "Auto"

A collection of code modules can be stored as a system library. The desired program can be achieved by concatenating different code modules from the system library with code generated by the user.

The user will specify to "Auto" what he would like to do, and "Auto" will advise him which methods are available for the solution, as well as their characteristics. The user will then choose the method he prefers, and "Auto" will generate the needed code.

The library currently contains the following modules.

1. Linear Search
2. Binary Search
3. Interchange Sort
4. Shell Sort
5. Sin(x) $0 \leq x \leq 2\pi$

$$\text{Sin}(x) = \sum_{n=0}^{m} \frac{(-1)^n x^{2n+1}}{(2n+1)!}$$

6. Cos(x) $-2\pi \leq x \leq 2\pi$

$$\text{Cos}(x) = \sum_{n=0}^{m} \frac{(-1)^n x^{2n}}{(2n)!}$$

7. $$\text{Ln}(x) = 2 \sum_{n=0}^{m} \left(\frac{1}{2n+1}\right)\left(\frac{x-1}{x+1}\right)^{2n+1}$$

8. $$\text{Exp}(x) = \sum_{n=0}^{m} \frac{x^n}{n!}$$

9. Arctan(x) for $|x| \leq |$

$$\tan^{-1}(x) = \sum_{n=0}^{m} \frac{(-1)^n x^{2n+1}}{2n+1}$$

10. Bessel Function of the First Kind and Zero Order

$$I_0(x) = \sum_{n=0}^{m} \frac{\left(-\frac{x^2}{4}\right)^n}{(n!)^2}$$

11. Bessel Function of the First Kind, Orders 0, 1 and 2

$$J_0(x) = \sum_{n=0}^{m} \frac{\left(-\frac{x^2}{4}\right)^n}{(n!)^2}$$

$$\frac{2}{x} J_1(x) = \sum_{n=0}^{m} \frac{\left(-\frac{x^2}{4}\right)^n}{(n!)^2 \cdot (n+1)}$$

$$J_2(x) = \frac{2}{x} J_1(x) - J_0(x)$$

12. Modified Bessel Function of the First Kind, Orders 0, 1 and 2

$$I_0(x) = \sum_{n=0}^{m} \frac{\left(\frac{x^2}{4}\right)^n}{(n!)^2}$$

$$\frac{2}{x} I_1(x) = \sum_{n=0}^{m} \frac{(\frac{x^2}{4})^n}{(n!)^2 \cdot (n+1)}$$

$$I_2(x) = -\frac{2}{x} I_1(x) + I_0(x)$$

13. Error Function

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{m} \frac{(-1)^n x^{2n+1}}{(n!)(2n+1)}$$

14. Fresnel Integral C(x)

$$C(x) = \int_0^x \cos(\frac{\pi}{2} t^2) dt = \sum_{n=0}^{m} \frac{(-1)^n (\frac{\pi}{2})^{2n} x^{4n+1}}{(2n!)(4n+1)}$$

15. Fresnel Integral $C_2(x)$

$$C_2(x) = \frac{1}{2\pi} \int_0^x \frac{\cos t}{t} dt = \frac{1}{2\pi} \sum_{n=0}^{m} \frac{(-1)^n x^{2n+1/2}}{(2n!)(2n+1/2)}$$

16. Fresnel Integral S(x)

$$S(x) = \int_0^x \sin(\frac{\pi}{2} t^2) dt = \frac{\pi}{2} \sum_{n=0}^{m} \frac{(-1)^n (\frac{\pi}{2})^{2n} x^{4n+3}}{[(2n+1)!](4n+3)}$$

17. *Sine Integral* $S_i(x)$

$$S_i(x) = \int_0^x \frac{\sin t}{t} dt = \sum_{n=0}^{m} \frac{(-1)^n x^{2n+1}}{[(2n+1)!](2n+1)}$$

18. Cosine Integral Cin(x)

$$\text{Cin}(x) = \int_0^x \frac{(1 - \cos t)}{t} dt = -\sum_{n=1}^{m} \frac{(-1)^n x^{2n}}{[(2n)!](2n)}$$

19. Dilogarithm f(x)

$$f(x) = -\int_1^x \frac{\text{Ln } t}{(t-1)} dt = \sum_{n=0}^{m} \frac{(-1)^n (x-1)^n}{n^2}$$

Note: m is so chosen that the magnitude of the m-th term of the power series is less than or equal to 10^{-6}, while the magnitude of the (m-1)-th term is greater than 10^{-6}.

The development of "AUTO-PROGRAMMING" will continue to concentrate mainly on increasing the size of the library. More mathematical programs, as well as some utility programs for data manipulations, will be developed in the near future.

Rome Air Development Center
F30602-74-C-0294

E. Lipschitz

SMALL SCALE TESTS

H. Ruston and M. L. Shooman

In order to verify the theoretical models four programs have been written by student programmers and careful records were kept on their debugging experiences. We will describe the assigned programs and the error reporting form. This data is presently being reduced and the resulting conclusions will be described in a future report.

A. The Programmers

The programmers were undergraduate students of sophomore junior standing, with high interest and ability in programming topics. Because of this selectivity we believe their product to be likely the one of equivalent to the one of programmers with intermediate experience.

Consequently, we consider the obtained test data to be representative of normal normal practice.

B. The Instructions

The programmers were made aware of the importance of maintaining careful and truthful records. They were also given the following specific instructions:

(1) The problems were to be analyzed and coded, with both analyses and coding times recorded.

(2) The resulting program was to be corrected of just the syntax errors. Their number, number of runs needed for their correction, and run times were to be recorded. All print-outs were to be saved and numbered.

(3) The programs were then presented to us (i.e., M. Shooman and H. Ruston). We planned to ask the program author and other members of the group to debug each copy independently, recording:

   a. Number of bugs and types found in each debugging shot

   b. Analysis time and computer time for each debugging shot

   c. History of removed bugs and generated bugs.

(4) If a programmer reached a blind alley he had to consult with us. He could not ask for other help, or abort the program.

(5) The program had to be constructed with the following constraints:

   a. To be structured

   b. To contain no impurities (as listed in Halstead[1])

   c. The main program to be the control structure with calls to modules (i.e., blocks or procedures)

   d. No module to exceed 50 lines.

C. The Four Programs

Three small problems (Problems 1, 2 and 3) and one medium size problem were generated for the small scale tests. The initial write-ups of the problems for the desired four programs follow.

Problem No. 1 - Minimum Salary Payroll Adjustment

1. Statement of the Payroll Adjustment

Glen Cove University, which employs 200 faculty members, has just signed a non-faculty union contract. All salaries of $16,000 or higher are to remain unchanged. Any faculty member who earns less than $16,000 per year is to receive a pay raise according to the following formula: He will receive $100 per year additional for each dependent (including himself), plus $50 per year for each year of employment. In no case may his new salary exceed $16,000 per year.

The personal data on all faculty members is stored on magnetic tape in the Business Office and includes present annual salary, number of dependents, date of hire and other information. The problem is to write a program which computes and prints out the list of faculty members along with their old and new salary.

Assume that the Business Office will give you a set of cards with the data for each person on one card. You should create your own test data; however, final testing of your program will be done on the actual card deck. Assume the following arrays will accept the input data in your program:

NAME (200)	:	contains a name of up to 30 characters
DEPEND (200)	:	contains number of dependents in two decimal digits
DHIRE (200)	:	contains a string of 10 characters, two digits for month, a blank, two digits for day, a blank, and four digits for year
PRESS SAL (200):		contains 5 digits with yearly salary in rounded dollars.

2. Approaches (write two programs, using each approach):

(a) Search list of names for those making less than $16,000 per year, compute new salary, check $16,000 limit, store new salary, print output

(b) Sort list by salary from lowest to highest, stop when $16,000 is exceeded, compute new salary, check $16,000 limit, store new salary, print output.

3. Language and Computer: PLAGO on Poly 360/65.

Problem No. 2 - Finding the Roots of a Cubic Equation

1. Statement of Problem

The polynomial equation $a_3x^3 + a_2x^2 + a_1x + a_0 = 0$ is to be solved for its three roots. A general solution is desired which will work for any finite real values of a_3, a_2, a_1 and a_0. The values of a_3, a_2, a_1, and a_0 are to be acquired as

floating (single precision) input data at the beginning of each run. Write your program with a loop so it reads any number of data cards and terminates on the last data card. Make up your own test data; however, it will be tested finally with supplied data cards.

2. Approaches

(a) The general formula (i.e., Cardano's formula) for the solution of a cubic is to be used to compute the roots

(b) An iterative solution for a single real root is to be obtained. Once the real root is removed, the quadratic formula is to be used to solve for the other two roots.

3. Language and Computer: PLAGO on Poly 360/65.

Problem No. 3 - Manipulation of a File of Research Reports

1. Statement of the Problem

At present the Reliability, Safety, and Software Engineering Group at the Polytechnic has about 1,000 reports, papers, books, journal proceedings in its library. Each item is entered on an index card in a file box. It is anticipated that the library may eventually grow to 10,000 items in the future. In the near future, two punched cards will be created for each of the items in the library and we wish to create a program to perform various searches, sorts, and listings.

Assume that each punched card contains 4 character fields. The first field is 30 characters wide and contains the author(s) name(s). The second field is 50 characters wide and contains the full or abbreviated title. (No important words in the title are to be abbreviated.) The second card contains field 3 which is 50 characters wide and contains key words (or abbreviated) key words in the item. The last field is 30 characters wide and contains the source (journal, issue, book publisher, proceedings, etc.) of the item.

The program must be able to perform the following tasks, and the selection (and possibly sequence) of tasks to be performed must be controlled by the first data card which will serve as a program control card. Provide a means of performing tasks on the same run.

(1) Read in a variable length stack of item cards and store them

(2) Alphabetize the data by first author

(3) Print out the list of items

(4) Create a list of key words (from field (3)), eliminate duplicates, alphabetize, and print out

(5) Search the author field for a given author's name, and print out the list of items he has written;

(6) Search the key word field for items which contain the "intersection" (AND) of one, two, or three inputed key words

(7) Provide the same search facility as (6) on words in the title field.

Programmer should create his own test cards, and final testing will be performed by a supplied deck of item cards.

2. Approaches

Programmer should provide his own approaches.

3. Language and Computer: PLAGO on Poly 360/65.

Problem No. 4 - Specifications for Ballot Counting Procedure

1. Terms

An election consists of sets of ballots to elect persons for various committees. Each set of ballots is called a committee election. There may be up to 10 committee elections for an election.

All ballots for a particular committee have the name of the committee punched in columns 61-80 and contain the names of the candidates to that committee. There are up to 25 nominees for each committee.

When a person votes for a candidate, an "11" punch (i.e., a minus sign) is punched in the ballot in the field consisting of columns 21-45 corresponding to candidates 1-25. Such a punch is a mark. A particular ballot may have more than one mark since a particular committee may have more than one vacant position (i.e., there may be more than one vote allowed to each voter on each ballot).

An election package consists of one ballot for each committee. Each eligible voter receives one and only one election package which he marks and returns for counting.

2. Specifications

Prior to counting the ballots, the program must read certain preliminary information concerning the election.

For each committee election the program must be informed as to:

(1) The name of the committee

(2) The number of candidates

(3) The number of marks permitted (e.g., if "vote for three" then three marks are allowed).

A rough flow chart is shown below.

DETERMINE VALIDITY

(1) Check election name (cols. 61-80). If name is found then

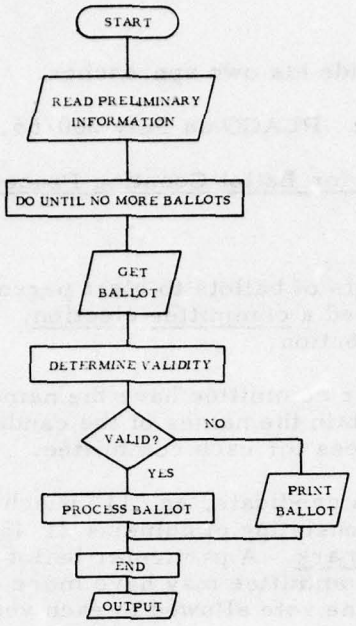
(2) Check all marks (cols. 21-45) to see if all are '-' or ' '. If ok then

(3) Check that no mark occurs after last candidate.

If ok then

(4) Count number of marks. If total is less than or equal to the number of marks permitted, then ballot is valid.

If any test fails then ballot is invalid.



PROCESS BALLOT

Consists of tallying the marks in some sort of array, probably 25 x 10.

OUTPUT

Printing tallies.

Typical output for 3 committees named PPC, SAB, TENURE

CANDIDATE NUMBER	PPC	SAB	TENURE
1	5	3	9
2	2	2	2
3	6	8	3
4	7	7	5
5	9	1	8
6	1	0	7
7		4	2
8		0	5
9		0	
10		8	

THERE WERE 50 VALID BALLOTS

THERE WERE 4 INVALID BALLOTS AND THESE ARE REPRINTED ON PREVIOUS PAGE

If there are no invalid ballots, then the last line can be left unprinted. The sample output has 10 rows of tallies since SAB has 10 candidates. Provision should be made for printing any number of rows from a minimum of 2 to a maximum number of candidates for any committee.

D. The Reporting Form

After several revisions the resulting reporting form has been selected.

POLYTECHNIC INSTITUTE OF NEW YORK
Department of EE/EP
Division of Computer Science
Safety, Reliability and Software Engineering Group
RADC Contract F30602-74-C-0294 (Softy)

INDIVIDUAL ERROR REPORTING FORM
(This must be completed for each non-syntax error.)

1. Identification
   (a) Programmer's Name_____
   (b) Program Title_____
   (c) Date_____
   (d) Form Number (for this program)_____
   (e) Description of the error (be precise)_____
   _____
   _____
   (f) Description of the correction (be precise)_____
   _____
   _____

2. Means of Detection - √only for Corrections (not New Reqs.)
   - More than one category may be √'ed

   ☐ a. Hand Processing
   ☐ b. Personal Communication
   ☐ c. Infinite Loop
   ☐ d. Interrupt Error (Code ___)
   ☐ e. Incorrect Output or Result
   ☐ f. Missing Output
   ☐ g. Other - Explain

3. Effort to Diagnose the Error - Do not include effort spent in initial detection
   a. No. of Runs to Diagnose____ Elapsed Computer Time (Minutes)_____
   b. Working Time to Diagnose Hours _____

4. Category of Change
   SOFTWARE CHANGE REQUIRED
   Nature of Change
   ☐ Documentation (Preface or Comments)
   ☐ Fix Instruction
   ☐ Change Constants
   ☐ Structural
   ☐ Algorithmic
   ☐ Other - Explain

Source of Bug

___
___ Bug essentially unrelated to previous corrections (i.e., usual case of bug just discovered)
___ Previous correction did not remove the believed error (i.e., improper or incomplete analysis)
___ New bug, introduced by a previous correction (i.e., bug generation through a correction)

Type of Bug

___
___ Misinterpretation of Specifications
___ Wrong Specifications
___ Incomplete Specifications
___ Incorrect Sequencing of Computations
___ Incorrect Input Data (Type and Quantity)
___ Incorrect Expressions
___ Incorrect Declaration
___ No Defense Against Invalid Data

___
___ Operating System
___ Support Software
___ Card Mispunched
___ Other - Explain

5. Difficulty of Correction

   a. No. of Runs to Correct ____ Elapsed Computer Time (Minutes) _____

   b. Working time to Debug: Days _____ Hours _____

   c. No. of Cards: Changed _____ Added _____ Deleted _____

6. Comments (Use Reverse Side and Additional Sheets if Necessary)

E. The Present Status and Planned Activities

Problems 1, 2 and 3 have been debugged by single programmers, that is, by just their authors. A set of test data has been selected for problem 1, and the several versions of the first program have been exercised with this test data successfully.

It is planned to perform the additional debugging with other programmers and to use the test data for the experimental small scale verification of our theoretical work.

Rome Air Development Center
F30602-74-C-0294

H. Ruston and M.L. Shooman

REFERENCE

1. M. Halstead, "Software Physics," Basic Principles IBM Research Report, RJ1582, IBM Research, Yorktown Heights, N.Y. (May 1975).

MICRO RELIABILITY MODELS

M. L. Shooman

Many previous software reliability prediction models by this author[1] and others[2] have concentrated on the bulk (macro) aspects of a program. This work involves a newly developed micro model[3] which is based on program structure.

It is assumed that the program has been written in structured or modular form so that decomposition into its constituent parts is simple. Further, we assume that via analysis of the program the decomposition can be related to several paths or other functional structures within the program.

The model is constructed based upon the frequencies with which each of the j paths are run, (f_j), the running time of each path, (t_j), and the probability of error along each path, (q_j)

Several methods of calculating or measuring the f_j, t_j and q_j parameters are suggested. In fact it is possible to use one technique (historical data) to produce crude estimates at the start of the design, and refine the estimates with more accurate values as the design progresses. Given the existence of such a model, we can consider the application of three important design techniques which are impossible with a macroscopic model:

(1) Apportionment of the software reliability (or mean time to failure) specification among the subsystems so each design team has their own goal to meet. *The apportionment* is obviously done so that the subsystem reliabilities combine to yield a system reliability which meets system specifications.

(2) If design proceeds either bottom up or top down, eventually there is a system integration phase where all parts are put together and tried out. The macroscopic models developed previously could not be applied before the system reached the integration stage. However, the new microscopic model proposed can be used to combine the results of the module development phases to predict a preliminary software reliability index before the system integration phase.

(3) The microscopic model is based upon measurements made on the software design. Such measurements and analyses performed on the software lead to a more disciplined design and provide insight into how the module performance relates to overall software system performance.

A. Micro Decomposition Model

The micro decomposition model which will be proposed in this section is based upon several assumptions. We first assume that the program has been designed using a structured or modular philosophy and as a result there emerges a natural structure of the program which can be described as consisting of a number of paths, cases, parts, modules, or subprograms. The decomposition focuses about this natural structure. In general we will primarily use the term paths from now on to designate the paths,

cases, parts, modules, subprograms, or any other important substructure. We also assume that the majority of the paths are independent of each other. (One could probably tolerate some type of dependence in the model if it were limited.)

The decomposition model will be developed from the probablistic viewpoint of relative-frequency. We will hypothesize a sequence of tests which either uncover a bug (failure) or run to completion without uncovering a bug (success). We begin our development of the model by defining the following variables and parameters:

$N \equiv$ The number of tests

$i \equiv$ The number of software paths (cases, parts, modules, etc.)

$t_i \equiv$ Time to run case i (if time is not deterministic we can substitute the mean value of t_i, i.e., $\overline{t}_i$)

$q_i \equiv$ Probability of error on each run of case i (the probability of no error $p_i = 1 - q_i$)

$f_i \equiv$ Frequency with which case i is run

$n_f \equiv$ Total number of failures in N tests

$H \equiv$ Total cumulative test time in hours.

Note that in the above set of definitions we have defined N as the number of tests. Thus, we are modeling actual or simulated operation by a succession of N tests (path traversals) of the system. We also assume the input data varies on each traversal. This is the reason why we have assigned a constant as the probability of encountering a bug on each run, q_i.

If there were no variation in input parameters, and three successive tests each traversed path j, then the probability of encountering an error on the first trial would be q_i. The conditional probability of encountering a bug on the second traversal of the same path with the same parameters is unity. Similarly, the probability on the same path with the same parameters the third time is also unity. Thus, the probability of a bug on three traversals of path j is $q_j \times 1 \times 1 = q_j$.

Since we have assumed a variation in parameters on each run in our model, each test is independent, then the probability of encountering one bug on three successive traversals of path j is given by the binomial distribution as[3]

$$P(1 \text{ error in three trials}) = \binom{3}{1} q_j^1 (1 - q_i)^2 \qquad (1)$$

Similarly, the expected number of occurrences in a probabilistic process governed by a binomial distribution is

$$\text{Number of Occurrences} = N_q \qquad (2)$$

where N is the number of trials and q the probability of occurrence.

B. Development of the Model

We can now compute the total number of failures n_f in N tests. The tests are distributed along each path such that Nf_1 tests traverse path 1, Nf_2 tests traverse path 2, etc. Thus, successive application of Eq. (2) to each of the i paths yields for the number of failures in N tests.

$$n_f = Nf_1q_1 + Nf_2q_2 + \cdots Nf_iq_i \tag{3}$$

We can now compute the system probability of failure on any one test run, q_o, by taking the ration of n_f/N as N approaches infinity

$$q_o \equiv \lim_{N\to\infty} \frac{n_f}{N} = \sum_{j=1}^{i} f_jq_j \tag{4}$$

Similarly we can compute the system failure rate, z_o, by first computing the total number of test hours. First we compute the total number of traversals of path i as Nf_i as was previously done. Out of these traversals the fraction p_i will be successful and will accumulate $N \times f_i \times p_i \times t_i$ hours of successful operation. If we assume that the time to failure distribution for the Nf_iq_i traversals which result in failure, is rectangular, then each trial which results in failure runs $t_i/2$ hours on the average before failure. Thus, the total test time accumulated in N runs is given by

$$H = Nf_1p_1t_1 + Nf_1q_1\frac{t_1}{2} + Nf_2p_2t_2 + Nf_2q_2\frac{t_2}{2}$$

$$+ \cdots Nf_ip_it_i + Nf_iq_i\frac{t_i}{2} = N\sum_{j=1}^{i} f_it_i(p_i + \frac{q_i}{2}) \tag{5}$$

Substitution for $p_i = 1 - q_i$ in Eq. (5) and simplification yields

$$H = N\sum_{j=1}^{i} f_it_i(1 - \frac{q_i}{2}) \tag{6}$$

We now compute the system failure rate z_o as

$$z_o = \lim_{N\to\infty} \frac{n_f}{H} \tag{7}$$

and substitution from Eqs. (3) and (6) into Eq. (7) yields in the limit

$$z_o = \frac{\sum_{j=1}^{i} f_jq_j}{\sum_{j=1}^{i} f_j(1 - \frac{q_j}{2})t_j} \tag{8}$$

C. Special Cases

We now wish to examine Eqs. (4) and (8) under special constraints. These are listed in Table I. Note that the units of z_o are clearly seen from Case 4 to be failures per hour, or just hr.$^{-1}$.

TABLE I. System probability of failure and failure rule for special cases.

Constraints $q_o \equiv \sum_{j=1}^{i} f_j q_j$		$z_o = \dfrac{q_o}{\sum_{j=1}^{i} f_j (1 - \frac{q_j}{2}) t_j}$
Case 1		
$q_j \ll 1$	$\sum_{j=1}^{i} f_j q_j$	$\dfrac{q_o}{\sum_{j=1}^{i} f_j t_j}$
Case 2		
$q_j \ll 1$		
$f_1 = f_2 = \cdots f_i = f = \frac{1}{i}$	$\frac{1}{i} \sum_{j=1}^{i} q_j$	$\dfrac{\sum_{j=1}^{i} q_j}{\sum_{j=1}^{i} t_j}$
Case 3		
$q_j \ll 1$		
$f_1 = f_2 = \cdots f_i = f = \frac{1}{i}$	q	$\dfrac{q_i}{\sum_{j=1}^{i} t_j}$
$q_1 = q_2 = \cdots q$		
Case 4		
$q_j \ll 1$		
$f_1 = f_2 = \cdots f_i = f = \frac{1}{i}$		
$q_1 = q_2 = \cdots q$	q	$\dfrac{q}{t}$
$t_1 = t_2 = \cdots t$		

D. Measurement of Parameters

In order to implement the model developed in the previous sections we must develop numerical values for the sets of parameters f_j, q_j and t_j. Of course, in keeping with the concept of structured programming and levels of structures within levels, one could merely state that we continue decomposition to lower levels until we end up with a new set f'_j, q'_j, and t'_j, parameters at a lower level. Clearly, the answer to the question how we could measure or estimate our parameters at a higher level is also, by and large, an answer to how we would do the measurement at a lower level.

The parameter sets f_j and t_j are related to the structure, size and complexity of the control structure and program modules. The determination of the f_j can be made by a study of the physical meaning of the paths and the distributions of input parameters which drive one along the program paths. If the program is complex, or there is really no information on input statistics, we can take one of two approaches. Assume f_j has a uniform distribution (see Case 2 of Table I) or insert counters in the various paths, and experimentally determine the f_j. The experimental approach requires that the program be in reasonably good shape so that a simulated test program can be run. Clearly, if a counter c_j is placed in each path such that it registers one count for each path traversal, and we run N tests then $f_j = c_j/N$.

The set of t_j parameters can also be either calculated or measured. If the program is written in assembly, machine or microprogramming code, one can estimate quite closely the run time of a sequence of code by summing the operating times of each instruction. If the program is complex, one can write an analysis program to read the code and perform the time analysis to determine t_j. In the case of a higher level language, (FORTRAN, PL/1, COBOL, etc.) the analysis is more complex, because each statement may expand into one to say ten machine language statements. Several approaches are possible. First of all, one can obtain a core dump of the machine language program and proceed as has been described. Another alternative is to insert a block of higher level code inside DO I = 1 to K loop. The loop is run for a particular value of K and the C.P.U. time of the computer recorded. The value of K is changed and another run and value of C.P.U. time is recorded. With about 3 values of C.P.U. time vs. K an accurate enough straight line or polynomial model of run time vs. K can be fixed to the data.* One can then use the formula to predict the run time of the actual

*It is necessary to take several measurements for two reasons. First of all there is program overhead which may vary from run to run depending on the operating system. (Also there is DO loop overhead.) Second, the recording of C.P.U. time is not accurate for short run times. To correct for DO loop overhead and also system overhead, one can perform the measurement with and without the code block in the loop and work with the difference.

code block by substituting the number of repetitions. (True value of K.) Of course, if the program and a simulation is available one can merely run several test runs for each path, record the times and use average values for each path.

The estimation of the q_j parameters is somewhat more difficult. During the early stages of design or development one can try and estimate q_i using historical data. One way to derive the q_j parameter is to obtain failure rate data on the program and equate it to z_o using the assumptions of Cases 3 or 4 in Table I, and solve for q_j. This process can be repeated as the program is written and better values for q_j determined. There is a possibility that one could calculate q_j from a more basic procedure. Knuth has shown that most FORTRAN statements are relatively simple and fall into one of several classes. If each of these classes also has a characteristic error rate, then by analysis of the q_j values for several examples, we should be able to derive characteristic values for the q_j parameters.

E. Conclusions

The models developed above allows one to decompose a program into a number of modules, paths, modes, or other functional entities. One can then compute an expression for the software failure rate in terms of probabilistic and deterministic parameters which can be estimated from historical data or determined by analysis or experiment. The model provides a clear cut procedure for relating the reliability of a large software system to the reliability of its constituent parts. The model is presently being applied to a number of modest size problems in order to obtain typical parameter values and validate the model.

Rome Air Development Center
F30602-74-C-0294

M.L. Shooman

REFERENCES

1. M. Shooman, "Probabilistic Models for Software Reliability Prediction," in Probabilistic Models for Software, Freiberger, Ed. (New York: Academic Press, 1972) pp. 485-502.
2. J. Jelinski and P.B. Moranda, "Software Reliability Research," in Probabilistic Models for Software, Freiberger, Ed. (New York: Academic Press, 1972).
3. M.L. Shooman, "Structural Models for Software Reliability Prediction," Proc. Second International Conf. on Software Engineering, San Francisco, California (October 1976).

NONLINEAR CONTROL OF LINEAR MULTIVARIABLE SYSTEMS

L. Shaw

It is well known that there are advantages in using nonlinear controllers for linear systems. Such controllers can achieve fast response to large errors, while responding slowly to small deviations which are often largely due to sensor noise contributions. While there have been some applications of this approach to single-input single-output systems, little is available in the way of general multivariable theory.

We have begun to develop an approach to this problem, starting from the theory of "linear system optimization with a prescribed degree of stability"[1] which constructs the optimal controller $u(\underline{x}, t)$ for the system

$$\dot{\underline{x}} = A\underline{x} + B\underline{u} \tag{1}$$

and the criterion

$$J = \int_0^\infty (\underline{x}'Q\underline{x} + \underline{u}'R\underline{u})e^{2\alpha t}dt \quad . \tag{2}$$

Here, A, B, Q, R, are constant matrices, (A, B) controllable, $R > 0$, $Q \geq 0$. Large values of the parameter $\alpha > 0$ force faster settling of an initial $\underline{x}(0) \neq 0$. It was shown in Ref. 2 that if α is sufficiently big, such that all eigenvalues of $(A + \alpha I)$ satisfy $Re\{\lambda_i(A+I)\} > 0$, then a solution exists in the limiting case $\lim_{Q \to 0} J = \bar{J}$. This case is interesting for two reasons. First, since α is a design parameter which provides for adjustment of settling time, one could argue that Q is unnecessary. (The restriction on the range of α values prevents a trivial solution of $\underline{u} = \underline{0}$.) Second, the limiting control law is in the linear feedback form

$$\underline{u}(t) = -R^{-1}B'P^{-1}(\alpha)\underline{x}(t) \tag{3}$$

where P is the unique positive definite solution of the Lyapunov equation

$$P(A+\alpha I)' + (A+\alpha I)P = BR^{-1}B' \quad . \tag{4}$$

This is in contrast to the necessity for solving a Riccati equation when $Q \neq 0$.

A nonlinear controller can be formed from the linear one above by making α a function $\alpha(\underline{x})$, such that α is bigger for "bigger" $\underline{x}$. To this end, we measure the size of $\underline{x}$ in terms of a Lyapunov function $V = \underline{x}'P^{-1}(\alpha)\underline{x}$. (It is straightforward to show that this is a Lyapunov function for the system in Eq. (1) using the controller in Eq. (3) for a fixed value of α.) We make $\alpha(\underline{x})$ an "increasing" function of $\underline{x}$ by the implicit definition

$$V(\underline{x}, \alpha) = F(\alpha) \tag{5}$$

where $f(\alpha)$ is some increasing function of α. $V(\underline{x}, \alpha)$ is also an increasing function since differentiation of Eq. (4) with respect to α shows $dP/d\alpha$ to satisfy

$$\frac{dP}{d\alpha}(A+I)' + (A+I)\frac{dP}{d\alpha} = 2P \quad , \tag{6}$$

and therefore to be negative definite so that

$$\frac{\partial V}{\partial \alpha} = \underline{x}' \frac{dP^{-1}}{d\alpha}\underline{x} = -\underline{x}' P^{-1}\frac{dP}{d\alpha}P^{-1}\underline{x} > 0 \quad . \tag{7}$$

It remains to show that a function $F(\alpha)$ can be found such that the constraint, Eq. (5), yields a unique α for each $\underline{x}$, and that the system using this $\alpha(\underline{x})$ in Eq. (3) is asymptotically stable. A class of suitable functions has been found to work in several examples, but the existence of such stable nonlinear controllers has not yet been demonstrated for all systems of the form Equation (1).

It is evident that

$$\dot{V} = -\underline{x}' P^{-1}[2\alpha P + BR^{-1}B']\ P^{-1}\underline{x} + \frac{\partial V}{\partial \alpha}\frac{\partial \alpha}{\partial t} \tag{8}$$

We assume that A has at least one stable mode so that the condition on eigenvalues of $(A+\alpha I)$ assures that $\alpha > 0$ and that the term in brackets in Eq. (8) is positive definite. For the other term, we note

$$\frac{\partial \alpha}{\partial t} = \frac{\partial \alpha}{\partial F}\frac{\partial F}{\partial t} = \frac{\partial \alpha}{\partial F}\frac{\partial V}{\partial t} \tag{9}$$

where the constraint in Eq. (5) has been invoked and the inverse function $\alpha(F)$ is well-defined due to the monotonicity of $F(\alpha)$. Thus Eq. (8) becomes

$$\dot{V}\left[1 + \frac{\partial V}{\partial \alpha}\Big/\frac{\partial F}{\partial \alpha}\right] = -\underline{x}' P^{-1}[2\alpha P + BR^{-1}B']\ P^{-1}\underline{x} \quad . \tag{10}$$

Using Eq. (7) it is clear that $\dot{V} < 0$ for all $\underline{x} \neq \underline{0}$, and $F(\alpha)$ is monotone increasing, if and only if

$$\frac{\partial F}{\partial \alpha} > \frac{\partial V}{\partial \alpha} \tag{11}$$

when $\alpha(\underline{x})$ is the solution of Eq. (5) for any $\underline{x}$. It is noteworthy that this condition also assures at most one solution to Equation (5).

A natural choice for $F(\alpha)$ is

$$F(\alpha) = [\underline{h}' P^{-1}(\alpha)\underline{h}]^k \tag{12}$$

for a fixed vector $\underline{h}$ and a constant k. We choose $\underline{h}$ such that $F(\alpha_{min}) = 0$, where $(-\alpha_{min})$ is the smallest (most negative) eigenvalue of A -- assumed temporarily to be

real for simplicity. The existence of such an $\underline{h}$ follows from multiplication of Eq. (6) on the left and right, respectively, by $\underline{h}' P^{-1}$ and its transpose to get

$$\underline{h}' \frac{dP^{-1}}{d\alpha} (P^{-1}AP + \alpha_{min} I)' \underline{h} + \underline{h}' (P^{-1}AP + \alpha_{min} I) \frac{dP^{-1}}{d\alpha} \underline{h} = -2\underline{h}' P^{-1} \underline{h} \tag{13}$$

Since $(P^{-1}AP + \alpha_{min} I)'$ has an eigenvalue of zero, we choose $\underline{h}$ to be the corresponding eigenvector and find $\underline{h}' P^{-1}(\alpha_{min})\underline{h} = 0$. (If the most negative eigenvalue of A is complex valued we use a complex α in Eq. (13), with P a function of $Re(\alpha)$, transposes replaced by conjugate-transposes, and $\underline{h}$ a complex valued vector.) This choice of $\underline{h}$ leads to F and V functions of the form shown in Figure 1. Since the slope of F can be increased

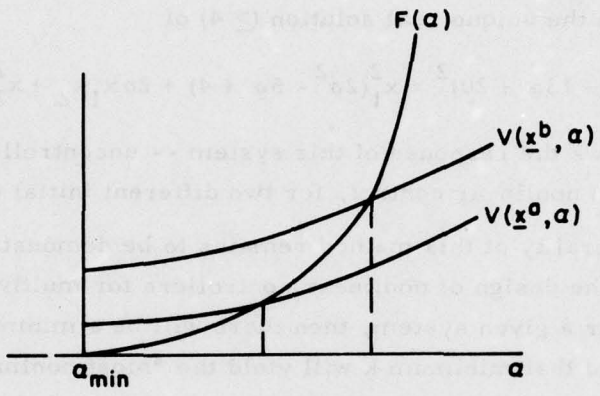


Fig. 1. Shapes of functions in Equation (5).

by increasing k, we can hope that a sufficiently large k will insure the existence of a solution to Eq. (5) as well as satisfaction of the stability condition, Equation (11).

When the F in Eq. (12) is used, the stability condition, Eq. (11), becomes

$$\underline{x}' \left[-\frac{dP^{-1}}{d\alpha} + P^{-1} k \underline{h}' \frac{dP^{-1}}{d\alpha} \underline{h} \Big/ (\underline{h}' P^{-1} \underline{h}) \right] \underline{x} > 0 \tag{14}$$

for all $\underline{x}$. Positive definiteness of the matrix in brackets is equivalent to that of

$$[-\Lambda + kI \, \underline{h}' \frac{dP^{-1}}{d\alpha} \underline{h} \Big/ (\underline{h}' P^{-1} \underline{h})] > 0 \tag{15}$$

where Λ is a diagonal matrix whose entries are the roots of the determinental equation

$$\left| \frac{dP}{d\alpha} + \lambda P \right| = 0 \quad . \tag{16}$$

Given a system defined by A, B, R, it is necessary to find a k such that the scalar

multiplying I in Eq. (15) is greater than max λ_i from Eq. (16) for all α. Some examples illustrate the utility of this approach.

(1) $\dot{x} = -ax + u; \quad a > 0; \quad R = 1$

Here any h can be used. With h = 1, the system is stable for $k > 1$; and when $k = 2$, $u = -(x^2)x$.

(2) $A = \begin{pmatrix} 0 & 1 \\ -4 & -5 \end{pmatrix}; \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}; \quad R = 1$

Here we find $\underline{h}' = c(1, -4)$ and the system is stable if $k > (2+\sqrt{2})/3 \approx 1.138$. With $k = 2$ and $c = 1$

$$u = -(4\alpha - 10)(\alpha x_1 + x_2)$$

where $\alpha(\underline{x})$ is the unique real solution (≥ 4) of

$$(4\alpha - 10)(2\alpha^2 - 13\alpha + 20)^2 = x_1^2(2\alpha^2 - 5\alpha + 4) + 2\alpha x_1 x_2 + x_2^2$$

Figure 2 shows the response of this system -- uncontrolled, with linear control, and with nonlinear control, for two different initial conditions.

Although the generality of this method remains to be demonstrated, it appears to be a first step toward the design of nonlinear controllers for multivariable systems. *If a suitable k exists for a given system,* then there will be a minimum suitable k. (See Equation (15).) Use of that minimum k will yield the "most nonlinear" system because larger k will make $F(\alpha)$ increase very steeply when $F(\alpha) > 1$, and will make all large values of $\underline{x}$ produce nearly the same α. Conversely, a large k will effectively lead to a smooth transition between two linear systems ($\alpha \approx \alpha_{min}$ producing the other linear system for small $\underline{x}$).

Implementation of the controllers described here requires solution of a polynomial equation at each sampling time. While the degree of this equation will generally increase as the system dimension increases, such calculations should be possible with present-day digital controllers. Indeed, feasibility of such a controller is one of the main advantages of digital versus analog controllers.

Use of a decreasing $F(\alpha)$ in Eq. (5) will clearly make $\dot{V} < 0$ in Equation (10). This corresponds to a kind of nonlinear system which puts a hard constraint on the control amplitude.

Arguments parallel to the ones given here yield similar results for receding horizon control[3] which always attempts to force $\underline{x}(t+T) = 0$. Nonlinear controls are produced there by making the horizon distance T a function $T(\underline{x})$ of the current state.

National Science Foundation L. Shaw
Centre National de La Recherche Scientifique

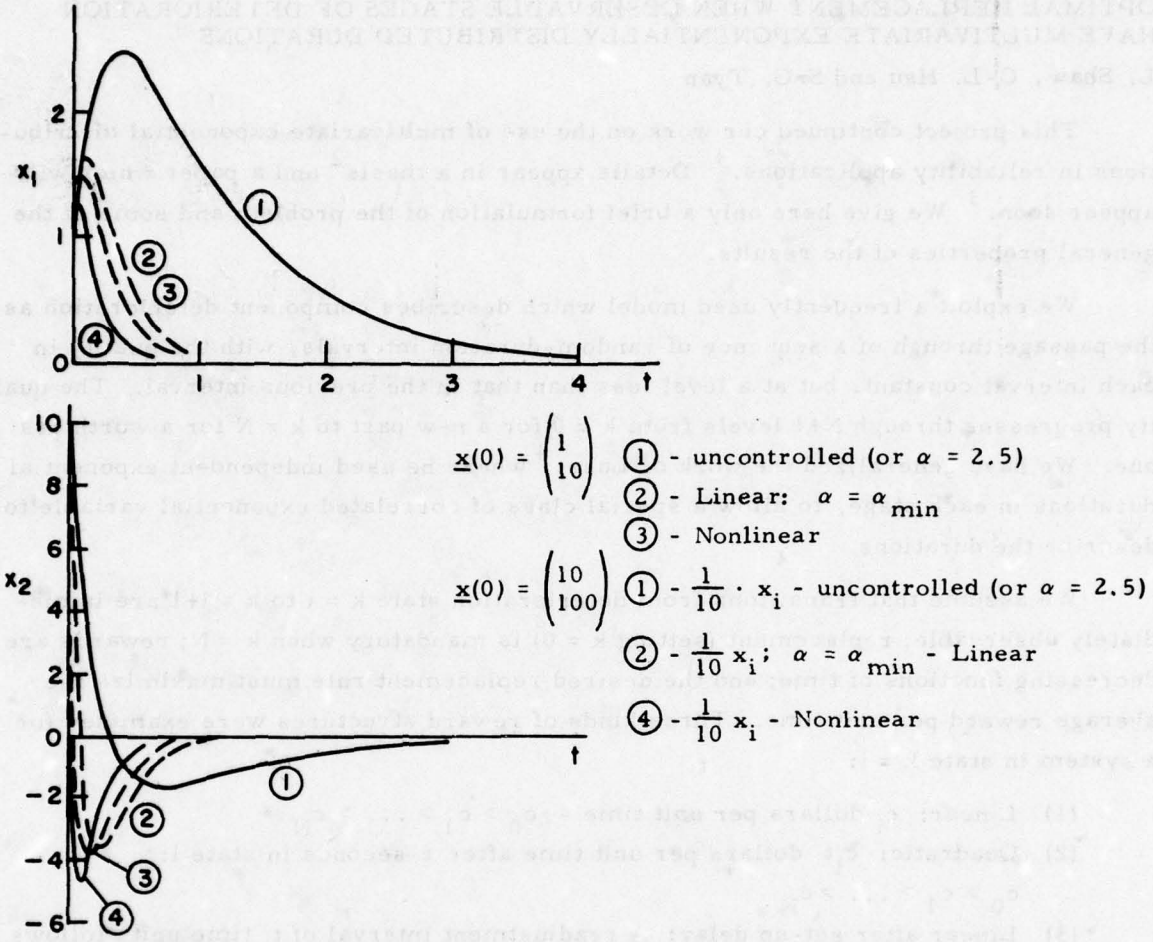


Fig. 2. Linear vs. nonlinear control.
Example No. 2.

REFERENCES

1. B.D.O. Anderson and J.B. Moore, "Linear System Optimization with a Prescribed Degree of Stability," Proc. IEE, Vol. 116, No. 12, pp. 2083-2087 (December 1969).

2. D. Sarlat and Y. Thomas, "Optimal Regulator with a Prescribed Degree of Stability: A Limiting Case," Electron. Lett., Vol. 11, No. 17, pp. 411-412 (August 1975).

3. Y. Thomas, D. Sarlat and L. Shaw, "A Receding Horizon Approach to the Synthesis of Nonlinear Multivariable Regulators," Electron. Lett. (1977).

OPTIMAL REPLACEMENT WHEN OBSERVABLE STAGES OF DETERIORATION HAVE MULTIVARIATE EXPONENTIALLY DISTRIBUTED DURATIONS

L. Shaw, C-L. Hsu and S. G. Tyan

This project continued our work on the use of multivariate exponential distributions in reliability applications.[1] Details appear in a thesis[2] and a paper which will appear soon.[3] We give here only a brief formulation of the problem and some of the general properties of the results.

We exploit a frequently used model which describes component deterioration as the passage through of a sequence of random-duration intervals, with the quality in each interval constant, but at a level less than that in the previous interval. The quality progresses through N+1 levels from k = 0 for a new part to k = N for a worthless one. We have generalized the work of Luss,[4] where he used independent exponential durations in each stage, to allow a special class of correlated exponential variable to describe the durations.

We assume that transitions from deterioration state k = i to k = i+1 are immediately observable; replacement (setting k = 0) is mandatory when k = N; rewards are decreasing functions of time; and the desired replacement rule must maximize the average reward per unit time. Three kinds of reward structures were examined for a system in state k = i:

(1) Linear: c_i dollars per unit time - $c_0 > c_1 > \ldots > c_N$
(2) Quadratic: $c_i t$ dollars per unit time after t seconds in state i: $c_0 > c_1 > \ldots > c_N$
(3) Linear after set-up delay: A readjustment interval of t_i time units follows the arrival in state k = i. The amount received is zero during that interval, but accrues at rate c_i afterwards.

The optimal rule for the linear case makes no use of the durations r_i in states k = i. It simply requires replacement when entering a state k^* whose value depends on the c_i and on the mean state durations.

In the other two cases, dynamic programming arguments show that the correlation between stage durations makes knowledge of these durations useful information for the optimal controller. There is a set of optimal decision thresholds r^*_{k-1} such that replacement is made on entering state k if $r_{k-1} < r^*_{k-1}$. A small r_{k-1} indicates that the current rate of deterioration is great. Furthermore, the optimal thresholds are ordered:

$$r^*_{k-1} \leq r^*_k$$

These results followed from a study of the properties of the conditional densities of the r_i sequence. In particular, that sequence preserves the Markov property of

the gaussian sequences which generate the correlated exponentially distributed variables, and the conditional variables are stochastically increasing in the sense that the conditional distribution function $F(r_k/r_{k-1})$ is an increasing function of r_{k-1} for every r_k. Moreover, that conditional density, expressible in terms of Bessel functions, is totally positive for all orders (TP_∞) and it defines conditional expectations which are convexity preserving.

Office of Naval Research
N00014-75-C-0858

L. Shaw, C-L. Hsu and S. G. Tyan

National Science Foundation
ENG 73-03803 A01

REFERENCES

1. L. Shaw and C-L. Hsu, "Reliability Applications of a Multivariate Exponential Distribution," Progress Report No. 41 to JSTAC, Polytechnic Institute of New York, Report No. R-452.41-76, pp. 502-506 (1976).

2. C-L. Hsu, "Reliability Applications of Multivariate Exponential Distributions," Ph. D. Dissertation (EE), Polytechnic Institute of New York (June 1977).

3. C-L. Hsu, L. Shaw, S. G. Tyan, "Optimal Replacement when Observable Stages of Deterioration Have Multivariate Exponentially Distributed Durations," Proc. Operations Research Society Meeting, Atlanta (November 1977).

4. H. Luss, "Maintenance Policies When Deterioration can be Observed by Inspections," Operations Research, Vol. 2, pp. 359 (1976).

ESTIMATION AND MODELLING OF NON-STATIONARY TIME SERIES

F. Kozin

A. Introduction

The object of this report is to present some ideas and recent results that we have obtained motivated by the study of non-stationary time series such as generated by strong motion earthquake records, and other geophysical phenomena.

We have approached this problem statistically. That is, given the data, fit a model. As yet, we have not based our studies upon the continuum mechanics of layered media, or other possible physical models or mechanisms. Naturally, for any modelling approach to be viable, it would have to be tested relative to known physical mechanisms.

Typical records of data that we have analyzed are shown in Figures 1 and 2. This data is taken from reports and tapes available from the Earthquake Engineering Research Laboratories of Cal Tech.[3]

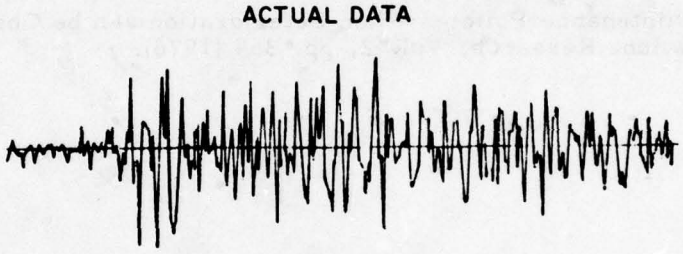


Fig. 1.

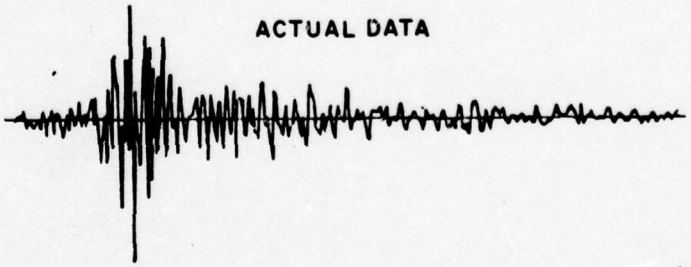


Fig. 2.

We have been concerned with acceleration data. However, it is clear that velocity or displacement data could be treated in the same fashion. Strong motion earthquake records are typically non-stationary in both amplitude as well as frequency

content. These characteristics are easily seen in Figures 1 and 2. Thus, traditional stationary statistical tools such as covariance and power spectral density analysis do not apply.

Purely on the basis of modelling of non-stationary time series, we divide the problem into three parts:

Postulate a class of models for the data.
Estimate the unknown parameters of the model from the data.
Assess the quality or fit of the model to the given data.

In this report, we shall treat each part separately indicating some of our pitfalls, as well as some of our positive results. Our initial studies of this problem can be found in References 1 and 2.

B. A Class of Models for Non-Stationary Time Series

The model that we postulate for the data is

$$y(k) + a_1(k-1)\,y(k-1) + \cdots + a_\ell(k-\ell)\,y(k-\ell) = g(k)\,u(k) \quad , \tag{1}$$

where y(k) represents the time series at "time" k, and $\{u(k)\}$ is a white Guassian sequence. The variance of u(k) will depend upon the time interval between data points. The coefficients $a_i(k)$, $i = 1, \ell$ as well as the coefficient g(k) are time varying and are to be obtained through the data.

Clearly the $a_i(k)$ cannot be completely arbitrary. Therefore, we parametrize the coefficients by assuming that they are of the form,

$$a_i(k) = \sum_{j=1}^{m} a_{ij} f_j(k) \quad , \quad i = 1, \cdots, \ell \quad , \tag{2}$$

where m is prechosen, and the $f_j(k)$, $i = 1, \cdots, m$ are a suitable family of known functions.

In order to give the $a_i(k)$ the freedom to fit a wide class of functions, we chose the $f_j(k)$ to be the orthogonal family of discrete Legendre polynomials,[4] defined as

$$f_j(k) = \sum_{s=0}^{j} (-1)^s \binom{j}{s} \binom{j+s}{s} \frac{k^{(s)}}{N^{(s)}} \quad , \tag{3}$$

$j = 0, 1, 2, \cdots, N$ where $k^{(s)} = k(k-1)\cdots(k-s+1)$, similarly for $N^{(s)}$, and N represents the number of data points. This family of polynomials is orthogonal on the interval $[0, N]$.

The function $\{g(k)\}$ plays the role of an envelope or amplitude modulation on the random noise input. It is also obtained from the data as described in the next section. In statistical terminology Eq. (1) represents a non-stationary autoregression model.

C. Parameter Estimation

We estimate the parameters a_{ij} by maximum likelihood techniques. Since no statistical theorems were available to determine the properties of maximum likelihood estimators for non-stationary autoregressive models, we proved a consistency result.[5] In its multi-variable form, the theorem is

Theorem - Consistency

Let

$$z(k) = A(k-1)\,z(k-1) + v(k) \quad , \tag{4}$$

where z, v are p-vectors, and

$$A(k) = A_0 + \sum_{i=1}^{m} a_i^0 A_i(k) \quad ,$$

where A_0, $A_1(k), \cdots, A_m(k)$ are given bounded p x p matrices. Finally, v(k) is a vector whose components are independent Gaussian white sequences, and whose non-singular covariance matrix is

$$E\{v(k)\,v^T(k)\} = \mathrm{diag}(g_i^2(k)) = G(k) \quad .$$

The constants a_i^0 represent the true, unknown, parameters that are to be estimated from the data.

The likelihood function for N data points is

$$L_N(a_1, \cdots, a_m) = -\sum_{k=1}^{N} \| z(k) - [A_0 + \sum_{i=1}^{m} a_i A_i(k-1)]\, z(k-1) \|_{G^{-1}} \tag{5}$$

If $\hat{a}(N)$ is the m-vector that minimizes $L_N(a_1, \cdots, a_m)$ for the N observations $z(1), \cdots, z(N)$ and if the matrix Q(N) satisfies for large N

$$\frac{1}{N} Q(N) > Q \text{ (Positive definite)}$$

where

$$Q_{ij} = \sum_{k=1}^{N} [A_i(k-1)\,y(k-1)]^T G^{-1}(k) [A_j(k-1)\,y(k-1)] \quad , \tag{6}$$

then

$$\lim_{N \uparrow \infty} \hat{a}(N) = a^0 \quad , \quad \text{where} \quad a^0 \quad \text{is}$$

the true parameter vector, with probability one.

AD-A050 265 POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 20/3
PROGRESS REPORT NO. 42, 15 SEPTEMBER 1976 - 14 SEPTEMBER 1977, --ETC(U)
NOV 77 A A OLINER F44620-74-C-0056
UNCLASSIFIED POLY-MRI-452.42-77 NL

6 OF 7
AD
A050 265



Having this result, we know that if one has sufficient data, the maximum likelihood estimates will converge to a set of limit values. This is a very important point that cannot be overlooked. When we first estimated these parameters in Refs. 1, 2, we applied recursive estimation techniques, in particular non-linear filtering, via the so-called Schwartz-Bass approximation.[6]

This approximation expands the non-linear terms about the (unknown) optimum estimate. Keeping terms up to and including second powers, a closed set of approximate filter equations are obtained which can be solved from the data to obtain approximate estimates. We found from our non-stationary model, that sometimes the estimates would converge and we obtained good results, other times the estimators would diverge and yield an overflow in our computer outputs. When we estimated the coefficients from data simulated by known non-stationary models, we found that if our initial guess was close to the known, true values, the Schwartz-Bass filter would yield estimates close to the true values. On the other hand, if the initial errors were relatively large, the computations would become unstable. We believe that this was due to the fact that since the recursive technique updates the estimate only one or two points ahead in time, it could not see that after an initial rise, say, in the amplitude characteristic of the data, it then decreases. Hence, the filter would read this as requiring an unstable system to fit the rise, which would then cause an overflow. We also mention in passing, that the so-called extended Kalman filter proved (in our experience) to be a poor estimator for coefficients even in the case of stationary models.

Since maximum likelihood estimates make use of the entire data, and since we know that these estimates are consistent by the theorem above, then we are always assured of stable computational results.

In our autoregressive model Eq. (1) the input u(k) is multiplied by the function g(k), which basically plays the role of the amplitude envelope of the original state variable y(k). To a certain extent the function g(k) separates the non-stationary amplitude characteristics of the data from the non-stationary frequency characteristics of the data. In order to see this, we need only recognize that the envelope g(k) is slowly varying relative to y(k). Thus, upon dividing the equality Eq. (1) by g(k) (note: $g(k) > 0$), the non-stationary amplitude characteristic is removed from the data. In our initial studies, we attempted to fit this envelope by linear combinations of two or three exponentials, but found that fitting the initial part of the envelope well, would lead to large errors in the later stages, and vice versa.

Hence, we were led to use another approach. We find that fitting the envelope of the data by cubic spline techniques leads to excellent results. That is, cubic polynomials are fit to various sections of the data, so that the points at which two polynomials

meet have equal derivatives. Thus, the envelopes become smooth curves. A typical envelope is shown in Figure 3.

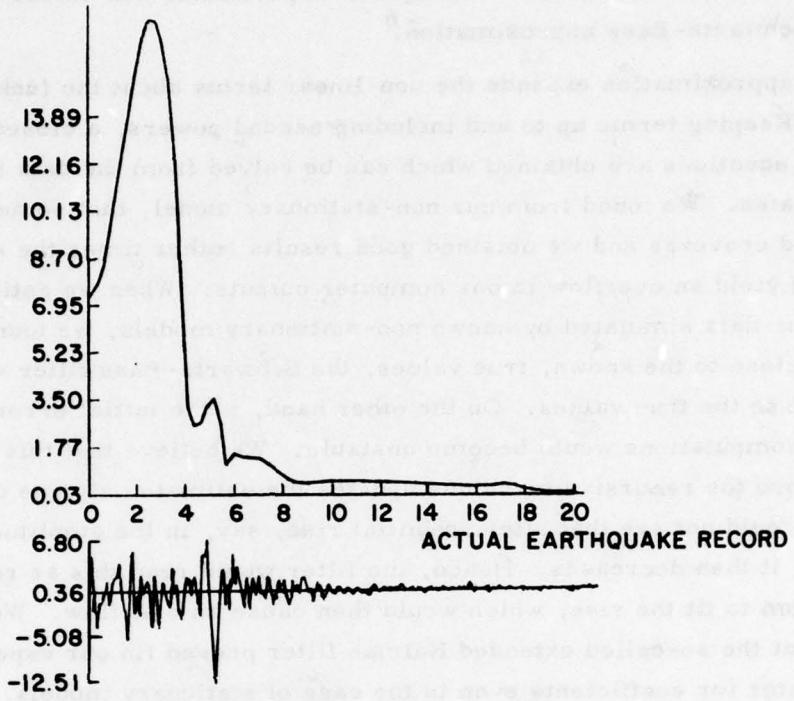


Fig. 3.

Our procedure therefore is as follows. Given a set of data, which typically from the Cal Tech reports is .02 seconds apart, our program first calculates the envelope function g(k) directly. Having the envelope function, and for a given value of ℓ and m in Eq. (2), maximum likelihood estimates for the unknown coefficients a_{ij} are determined for the given g(k).

In this way a model of the form Eq. (1) is fitted to the given non-stationary time series.

D. Quality of the Model

The natural question for any model fitting procedure is, "Is the model good"?

We can certainly look qualitatively at the model by generating records from Eq. (1) with the calculated g(k) and the estimated a_{ij}'s. This simulation simply requires that white Gaussian sequences $\{u_k\}$ be generated with variance proportional to .02, and used as an input to generate the y(k)'s in Equation (1).

Typical simulated records corresponding to Figs. 1 and 2 are shown in Figs. 4 and 5 respectively. Qualitatively, they look reasonably good. But the more important question is how shall we judge this model relative to other possible models?

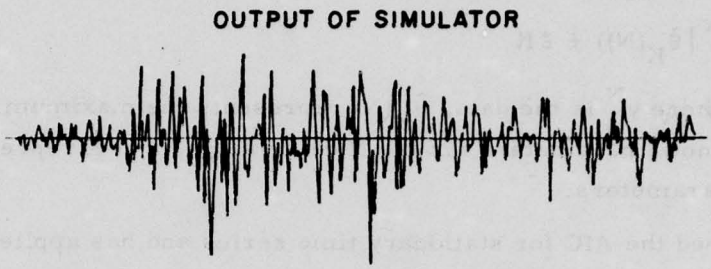


Fig. 4.

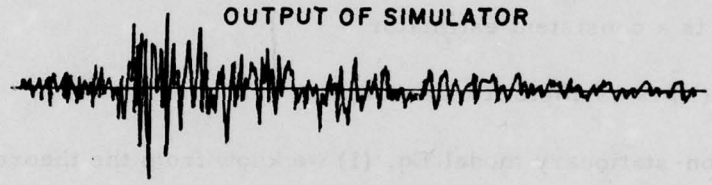


Fig. 5.

In our attempts to apply non-linear filtering methods in Refs. 1 and 2, we studied the quality of the model by studying the so-called residuals. The residual is the numerical difference between the estimated data value (from the model) and the actual data value. Theory tells us that the residuals should become a white noise sequence. Hence, we applied statistical tests to ascertain the whiteness of the residuals. Parameter estimates were accepted if the residuals passed several "whiteness" tests at the 95% confidence levels.

This approach appeared to be adequate in lieu of convergence theorems, but it is difficult to judge the relative quality of a given model in this way.

Fortunately, there is a recent approach to fitting autoregressive models due to Akaike[7] that appears to possess the potential for being a very useful tool.

Starting with the Kullback-Liebler mean information

$$E\left\{\log \frac{P(X|\theta)}{P(X|\hat{\theta})}\right\}$$

for discriminating between two conditional densities $p(X \mid \theta)$ and $p(X \mid \hat{\theta})$, Akaike derived the information criterion (AIC).

Akaike Information Criterion

Choose the model such that

$$-2 \log p(y^N \mid \hat{\theta}_K(N)) + 2K \tag{7}$$

is minimized, where y^N is the data, $\hat{\theta}_K(N)$ represents the maximum likelihood estimates of the unknown parameters for the N data values, and K represents the number of independent parameters.

He developed the AIC for stationary time series and has applied the criterion to a number of statistical problems including estimating the orders of auto-regressive moving-average models with excellent success.

In order to derive the criterion, Akaike requires two conditions

$$\left.\begin{array}{l} \text{(i)} \quad \hat{\theta}_K(N) \text{ is a consistent estimator} \\ \text{(ii)} \quad \sqrt{N}\,\hat{\theta}_K(N) \text{ is asymptotically normal.} \end{array}\right\} \tag{8}$$

For our non-stationary model Eq. (1) we know from the theorem of Section C that Eq. (8)(i) holds. Fortunately, we are able to establish Eq. (8)(ii) for our model also, via a Martingale Central Limit theorem.[8]

Thus, the AIC criterion holds for our non-stationary models as well.

We have applied the AIC criterion to study the best choices of ℓ, the order of the model, as well as the choice of the non-stationary coefficients.

To study the order of the model, the procedure is as follows. We first calculate the envelope function g(k). This remains fixed throughout all succeeding computations.

For each assumed value of ℓ, determine the maximum likelihood estimates for the a_i's. We note that for a given number m, of orthogonal polynomials used in the coefficients, there will be a maximum of $\ell \cdot m$ coefficients to estimate. Thus ℓm plays the role of K in Equation (7).

For each ℓ, and each set of computed estimates $\hat{a}_{ij}$ for that order ℓ, the expression Eq. (7) is calculated. That integer ℓ for which Eq. (7) takes its minimum value is chosen as the best order fit to the data.

We studied the best order fit of autoregressive models of the form,

$$y(k) + a_1 y(k-1) + \cdots + a_{\ell-1} y(k-\ell+1) + a_\ell(k-\ell)\, y(k-\ell) = g(k)\, u(k) \quad , \tag{9}$$

where $a_1, \cdots, a_{\ell-1}$ are unknown constants to be estimated and $a_\ell(k)$ is a time varying term of the form Eq. (2) for m = 5.

The results are:

IIB028-N88W

ℓ	AIC (ℓ)
2	3.4210×10^3
3	2.8856×10^3
4	2.9244×10^3

min $\ell = 3$

IIB031-S69

ℓ	AIC (ℓ)
2	2.2142×10^3
3	2.2702×10^3
4	2.2460×10^3

min $\ell = 2$

Many more computer experiments are being performed via the AIC and a number of interesting points are evolving. For example, records with a short strong motion period seem to be fit best by second order models. If the strong motion period is longer, third or fourth order models appear best. If the unknown coefficients are all assumed to be constant (stationary case), the best fit may be tenth order or higher.

Further studies are being performed to ascertain the significance of this approach to studying non-stationary time series.

National Science Foundation
GI 43095

F. Kozin

REFERENCES

1. R. Gran and F. Kozin, "*Nonlinear Filtering Applied to the Modelling of Earthquake* Data," Proc. Symp. on Non-Linear Estimation and Its Application, San Diego, California (September 1973).
2. F. Kozin and R. Gran, "An Approach to the Analysis and Modelling of Earthquake Data," in Stochastic Problems in Mechanics, Mech. Study (Waterloo, Ontario, Canada: University of Waterloo Press, 1974), No. 10, pp. 193-204.
3. "Strong Motion Earthquake Accelerograms," Tech. Rept., Earthquake Engineering Research Labs., Calif. Inst. of Tech., Vol. I (1970), Vol. II (1973).
4. W.E. Milne, "Numerical Calculus," (Princeton, N.J.: Princeton University Press, 1949).
5. T.S. Lee and F. Kozin, "Consistency of Maximum Likelihood Estimation for a Class of Non-Stationary Models," Proc. 9th Hawaiian Conf. on System Science, Honolulu, Hawaii (January 1976).
6. R. Bass, et al., "Optimal Multichannel Non-Linear Filtering," J. Math. Analy. Appls., Vol. 16, p. 152 (1966).
7. H. Akaike, "A New Look at Statistical Model Identification," IEEE Trans. Auto. Control, Vol. AC-19, pp. 716-723 (1974).
8. B.M. Brown and G.K. Eagelson, "Martingale Convergence to Infinity Divisible Laws with Finite Variances," Trans. Am. Math. Soc., Vol. 162, pp. 449-453 (1971).

RESTORATION OF IMPULSE TRAIN SIGNALS FROM SMOOTHED DATA

A. Papoulis and C. Chamzas

A central problem in signal analysis is the determination of a signal f(t) in terms of a band of its spectrum

$$F_\sigma(\omega) = F(\omega)\, P_\sigma(\omega) \tag{1}$$

where

$$F(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t)\, e^{j\omega t}\, dt \tag{2}$$

and

$$P_\sigma(\omega) \begin{cases} 1 & |\omega| < \sigma \\ 0 & |\omega| > \sigma \end{cases} . \tag{3}$$

Since only a smoothed version is known, f(t) can not be determined exactly. If the $f_o(t)$ is used to estimate f(t) then an error $f(t) - f_o(t)$ results. Generally two methods are used for the reduction of this error. Windows and extrapolation.

Windows: The known band of the signal is multiplied by a function $W(\omega)$ and the inverse Fourier transform $f_w(t)$ of the resulting product

$$F_w(\omega) = F_\sigma(\omega)\, W(\omega) \tag{4}$$

is used as the estimate of f(t). The function $W(\omega)$ is so chosen as to reduce in some sense the resulting error $f(t) - f_w(t)$. The main advantage of this method is its computational simplicity. However, it replaces the unknown part of $F(\omega)$ with zero. The method of windows has been extensively treated.[1]

Extrapolation: The known band $F_\sigma(\omega)$ is extrapolated beyond $|\omega| > \sigma$ and the inverse $f_e(t)$ of the so formed spectrum, $F_e(\omega)$ is used as the estimator for f(t). The nature of the extrapolation depends on various apriori assumptions and it is in general arbitrary. It is however no more arbitrary than the assumption that $F(\omega) = 0$ for $|\omega| > \sigma$ implicit in the method of windows. A special form of extrapolation technique is the method of maximum entropy. It is based on the assumption that the entropy of f(t) is maximum subject to the constraint that its Fourier transform $F(\omega)$ is given for $|\omega| < \sigma$ (see Reference 2). In Ref. 3 a new method on extrapolation is presented based on the assumption that f(t) is time-limited, i.e.,

$$f(t) = 0 \quad \text{for} \quad |t| > \tau .$$

In the next section we present an iteration scheme, based on Ref. 3, where the extrapolation of $F_\sigma(\omega)$ depends on the assumption that f(t) consists of impulses, i.e.,

$$f(t) = \sum_{i=1}^{k} a_i \, \delta(t - t_i) \quad . \tag{6}$$

We note that from Eq. (6) and Eq. (1), in the discrete case, a system of nonlinear equations for a_i and t_i can be formed. However the solution is computationally complex and noise sensitive.

A. Description of the Iteration

We are given the band $F_\sigma(\omega)$ of $F(\omega)$ and we wish to determine f(t) under the assumption of Eq. (6)

$$f(t) = \sum_{i=1} a_i \, \delta(t - t_i) \quad .$$

We use the following iteration. (Figure 1.) We start with $F_o(\omega) = F_\sigma(\omega)$ then if $G_n(\omega)$ is the estimator for $F(\omega)$ at the n-th step of the iteration we form the spectrum

$$F_n(\omega) = G_n(\omega) \, [1 - P_\sigma(\omega)] + F_\sigma(\omega) \tag{7}$$

Thus $F_{n+1}(\omega)$ is a better estimator since it is obtained by replacing the band of $G_n(\omega)$ from $-\sigma$ to σ with the known band of $F(\omega)$.

Then we compute the inverse $f_n(t)$ of $f_n(\omega)$

$$f_n(t) \longleftrightarrow F_n(\omega) \tag{8}$$

and we form the function

$$g_{n+1}(t) = \begin{cases} f_n(t) & \text{if} \quad |f_n(t)| \geq C_{n+1} \\ 0 & \text{if} \quad |f_n(t)| < C_{n+1} \end{cases} \tag{9}$$

by setting the parts of $f_n(t)$ which are less than C_{n+1} equal to zero. The truncation level C_n is defined as

$$C_{n+1} = \begin{cases} K_n & \text{if} \quad K_n < C_{max} \\ C_{max} & \text{if} \quad K_n \geq C_{max} \end{cases} \tag{10}$$

where K_n is the minimum value of $|g_n(t)|$ excluding the zero. The use of the truncation level C_n utilizes the additional information that f(t) is a signal of the form

$$f(t) = \sum_{i=1} a_i \, \delta(t - t_i) \quad .$$

The choice of C_o is not critical as long as the remaining non-zero part of $g_1(t)$ contains all the pulses locations t_i. The choice of C_{max} depends on our apriori knowledge of a_i and it must be less than the minimum expected value of a_i, that is

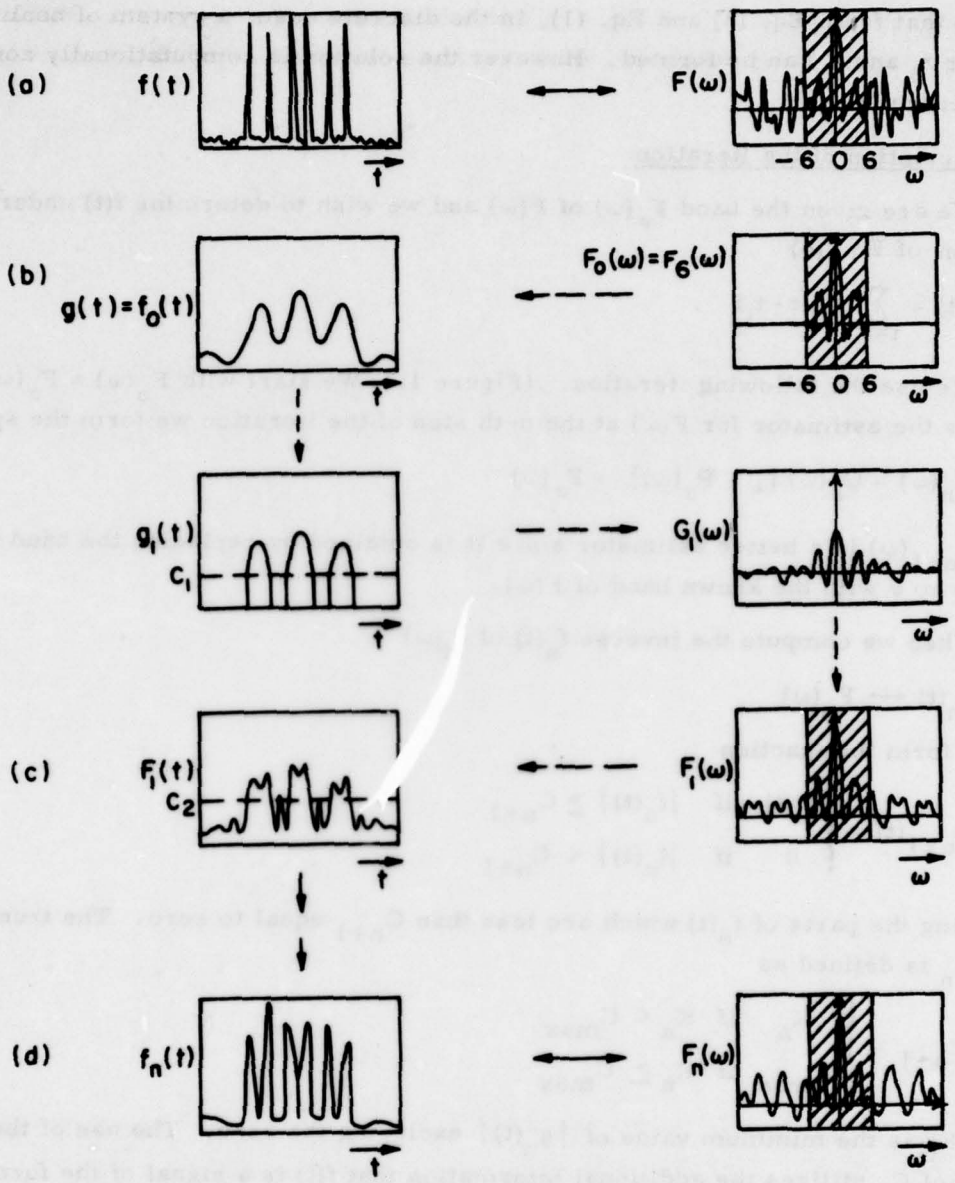


Fig. 1. Schematic of the iteration. Dotted arrows indicate iteration steps.
(a) Unknown signal f(t) and its Fourier transform (f(t) is not recognizable from g(t)).
(b) Obtained signal g(t).
(c) Result of the first iteration $f_1(t)$. (The signal starts to be recognizable.)
(d) Result of the 12th iteration $f_{12}(t)$.

$$C_{max} < \min(a_i) \tag{11}$$

As a result of using C_n, at each step of the iteration, $f_n(t)$ is concentrated closer to t_i and the length of the truncated region becomes smaller and smaller. At the end of the iteration when the pulses have been recovered the remaining non-zero part of $g_{n+1}(t)$ in Eq. (9) consists only of the points t_i.

B. Extension

If the noise is not white but colored, a proper window can be inserted in the algorithm weighting properly the given part of $F(\omega)$. Then instead of Eq. (7) we have

$$F_n(\omega) = G_n(\omega)[\,1 - W(\omega)\,] + F_\sigma(\omega)\, W(\omega) \quad \text{with} \quad 0 < W(\omega) < 1 \tag{12}$$

Also a constant A_n can be inserted in the algorithm to speed up the convergence. Then

$$F_n(\omega) = A_n G_n(\omega)[\,1 - W(\omega)\,] + F_\sigma(\omega)\, W(\omega) \tag{13}$$

where A_n is chosen from the minimization of

$$\int_{-\infty}^{\infty} |F(\omega) - A_n G_n(\omega)|^2 \left[1 - W(\omega)\right] d\omega = \min$$

Applications of the presented algorithm include detection of double or multiple stars, clutter elimination, determination of hidden periodicities and estimation of the location of impulse trains from smoothed data. Extension of the method in two-dimensional problems include sharp edge image restoration and line enhancement. Examples are given below indicating some of the applications.

Example 1: The first example is from the estimation of the location of impulse trains from smoothed data.

The unknown signal consists of six pulses with different amplitudes located as follows. (Δt is the sampling interval.)

Location	$\frac{t_i}{\Delta z}$	36	41	48	52	59	64
Amplitude	a_i	0.9	0.8	1.00	1.00	0.8	0.9

and

$$f(t) = \sum a_i\, \delta(t - t_i) + \underline{n}(t) \tag{14}$$

where n(t) is white noise uniformly distributed in [-0.05, 0.05]. Figures 2(a) and (b) show f(t) and the amplitude $|F(\omega)|$ of its Fourier transform. The given part of the $F(\omega)$ is shown in Fig. 3(b) and its inverse signal $f_o(t)$ in Figure 3(a). Notice that the six

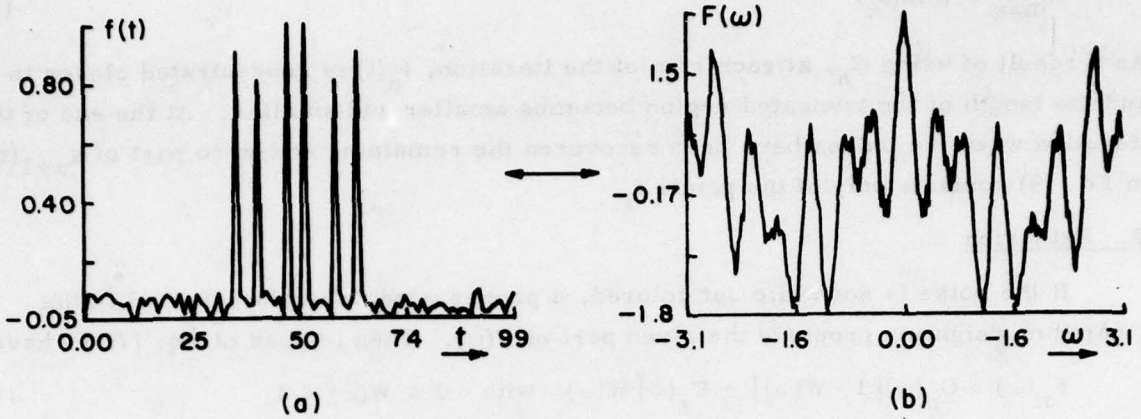


Fig. 2. The unknown impulse train and its Fourier transform.

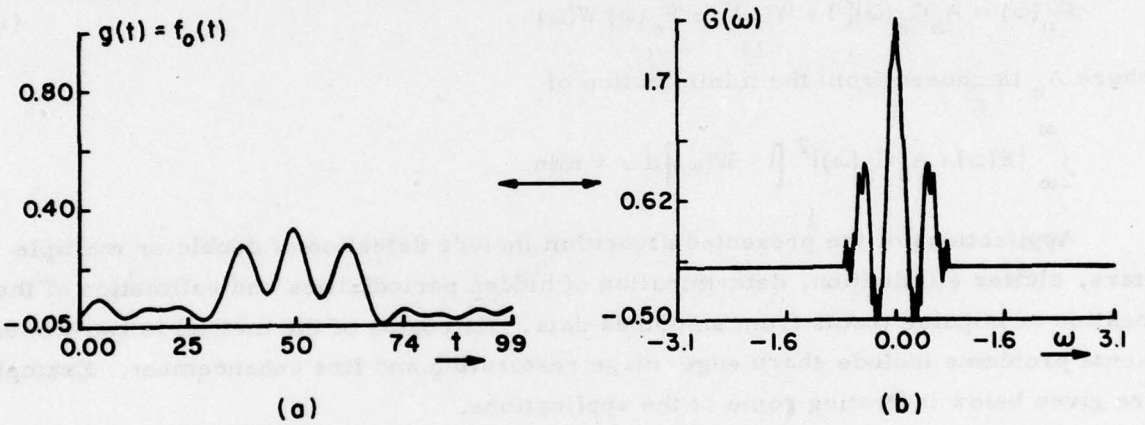


Fig. 3. The obtained signal g(t) and its Fourier transform.

pulses are not recognizable in $f_o(t)$. In Fig. 4 the 0-th, 12-th and 100-th steps of the iteration are shown. We observe that from the 12-th step the different pulses are recognizable and at the 100-th step they have been recovered completely.

Example 2: As a second example we will consider the problem of clutter elimination.

In this case the signal is

$$g(t) = f(t) + \underline{n}(t) \tag{15}$$

where $\underline{n}(t)$ is a uniformly distributed low frequency noise (clouds over a picture for example). In this case we intentionally perform a filtering to remove the unreliable frequencies. However, the filter also distorts our signal and some times makes it unrecognizable. In the present example, f(t) is shown in Fig. 5 with its Fourier transform pair $F(\omega)$. Figure 6 shows g(t) where a high level band-limited noise has been

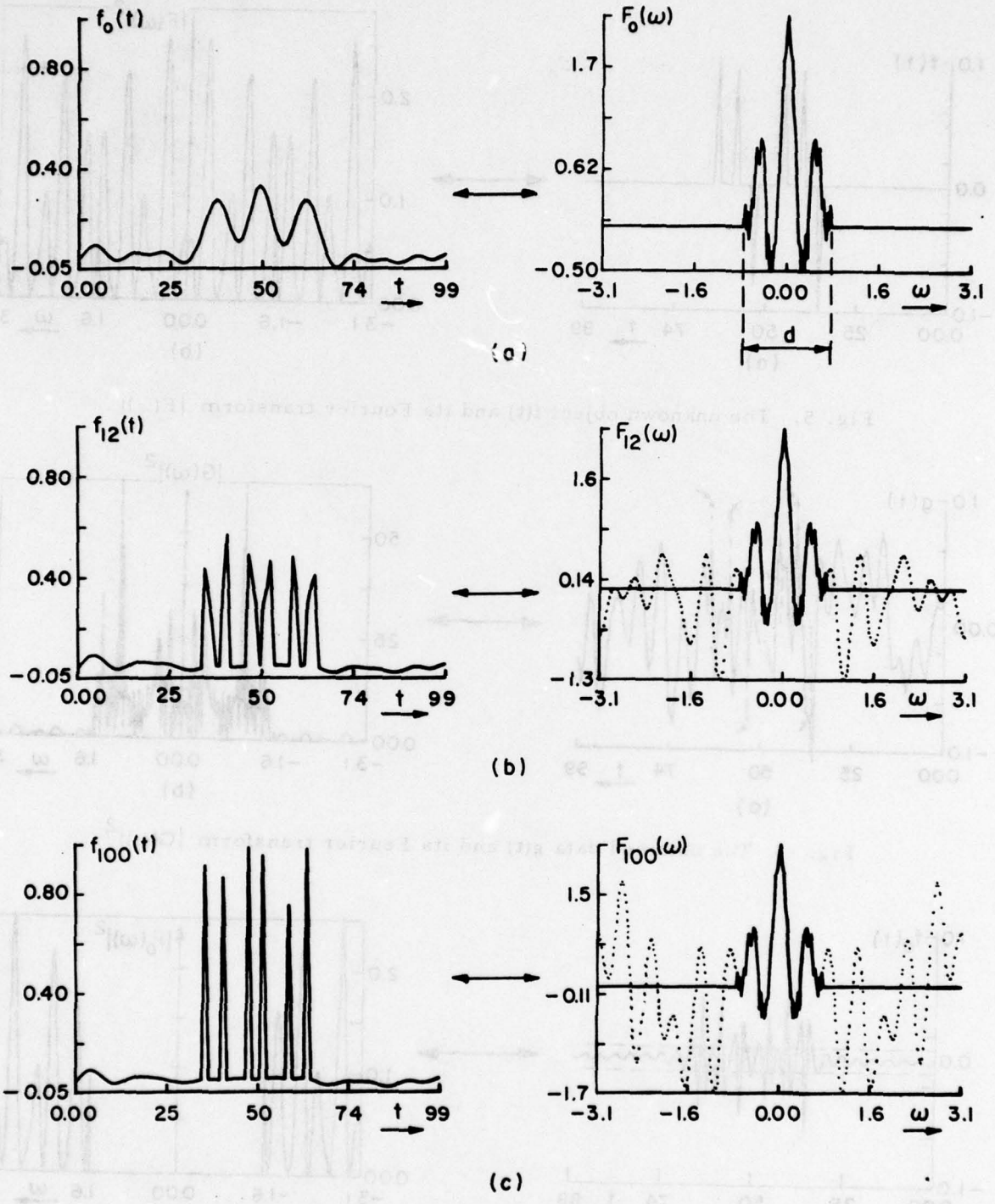


Fig. 4. (a) Given signal and its Fourier transform.
(b) The result of the 12-th iteration.
(c) The result of the 100-th iteration.

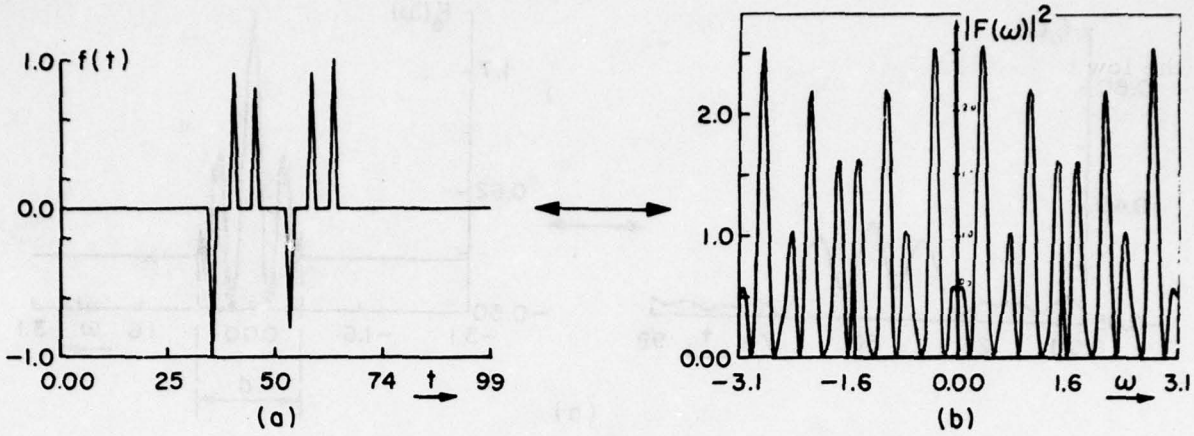


Fig. 5. The unknown object f(t) and its Fourier transform $|F(\omega)|$.

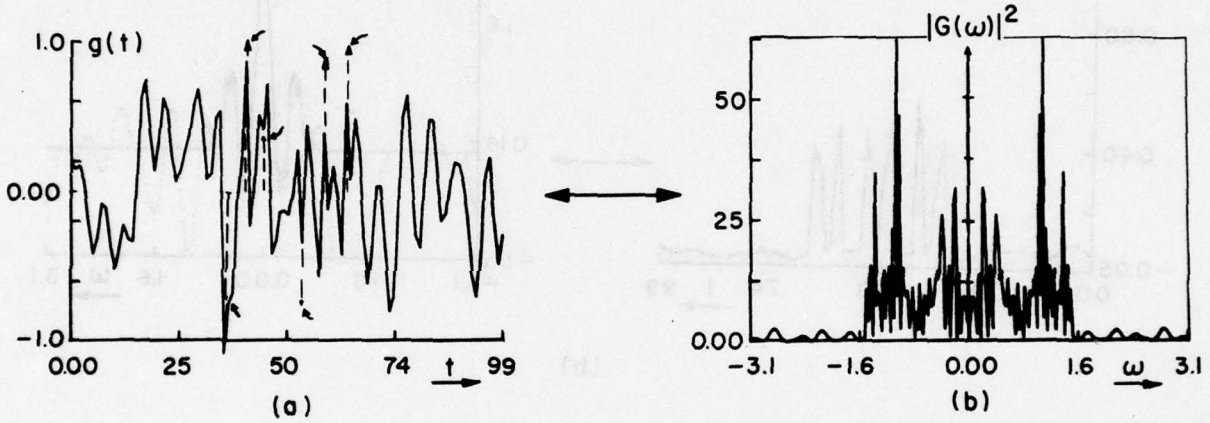


Fig. 6. The obtained data g(t) and its Fourier transform $|G(\omega)|^2$.

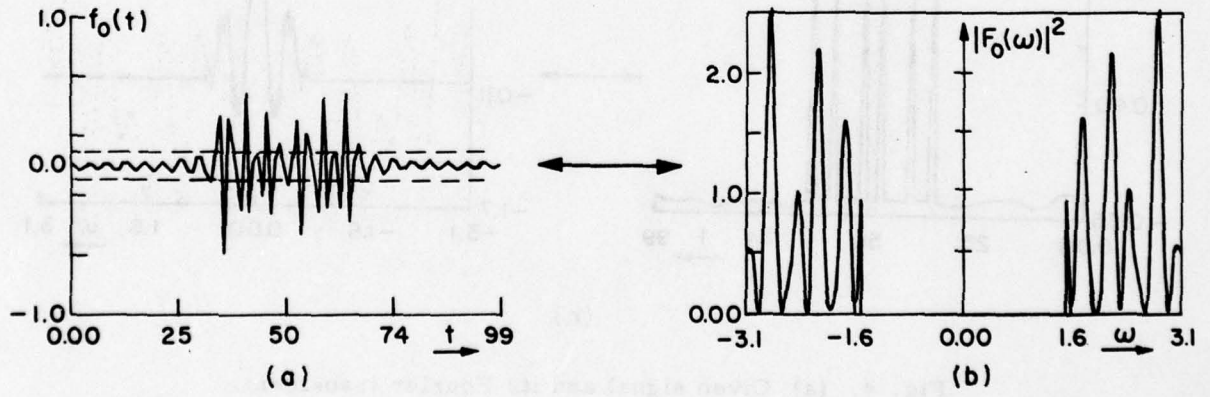


Fig. 7. The output of the high-pass filter.

added. (Its band is from -6 to 6.) The signal is entirely unrecognizable. Filtering the low band by a high-pass filter we remove the noise but again, as it appears in Fig. 7(a), the signal is unrecognizable since its low frequencies are missing (see Figure 7(b). We apply the algorithm for restoring these missing low frequencies. The iteration uses the formula Eq. (12) with

$$W(\omega) = 1 - P_{\sigma}(\omega) \tag{16}$$

The result of the iteration is shown in Figs. 8(a), (b) and (c) for the 0-th, 3-rd and 10-th steps. The signal again has been restored completely. Therefore by using the above algorithm we removed the filter effects.

Example 3: Hidden periodicities.

In a number of applications, the signal f(t) is a sum of sine waves of unknown amplitudes a_i and frequencies ω_i (sun spot variations for example). The transform $F(\omega)$ of f(t) consists of lines (impulses) at $\omega = \omega_i$, however since only the segment

$$f_T(t) = \begin{cases} f(t) & |t| < T \\ 0 & |t| > T \end{cases}$$

of f(t) is available, their locations can not be determined in terms of the transform $F_T(\omega)$ of $f_T(t)$. Interchanging t and ω on the above problem we are back in the problem of detecting an impulse train from its smoothed data. In this example we consider the case of determining the frequency and the amplitude of a sinusoid having only a part of its period. The given part is about the half period of an 1 Hz period and it is distorted by 10% white noise. In Fig. 9 we have 11 samples of the given part. The algorithm is applied to this sinusoid using a fast Fourier transform with 256 points. Figure 10 shows how the spectrum line of the sinusoid is emerging during the algorithm. Its location is exactly on the 1 Hz point of the spectrum. This result is more accurate than the one obtained in Ref. 4 through the maximum entropy method.

Advanced Research Projects Agency
N00014-76-C-0144

A. Papoulis and C. Chamzas

REFERENCES

1. R.B. Blackman and J.W. Tukey, "The Measurement of Power Spectra," (New York: Dover Publications, 1953).
2. R.T. Lacoss, "Data Adaptive Spectral Analysis Methods," Geophys., Vol. 86, pp. 661-675 (1971).
3. A. Papoulis, "A New Algorithm in Spectral Analysis and Band-Limited Extrapolation," IEEE Trans. Circuits Syst., Vol. CAS-22, No. 3, pp. 735-742 (September 1975).
4. T.J. Urtych, " Maximum Entropy Power Spectrum of Truncated Sinusoids," J. Geophys. Res., Vol. 77, No. 8, pp. 1336-1400 (Match 1972).

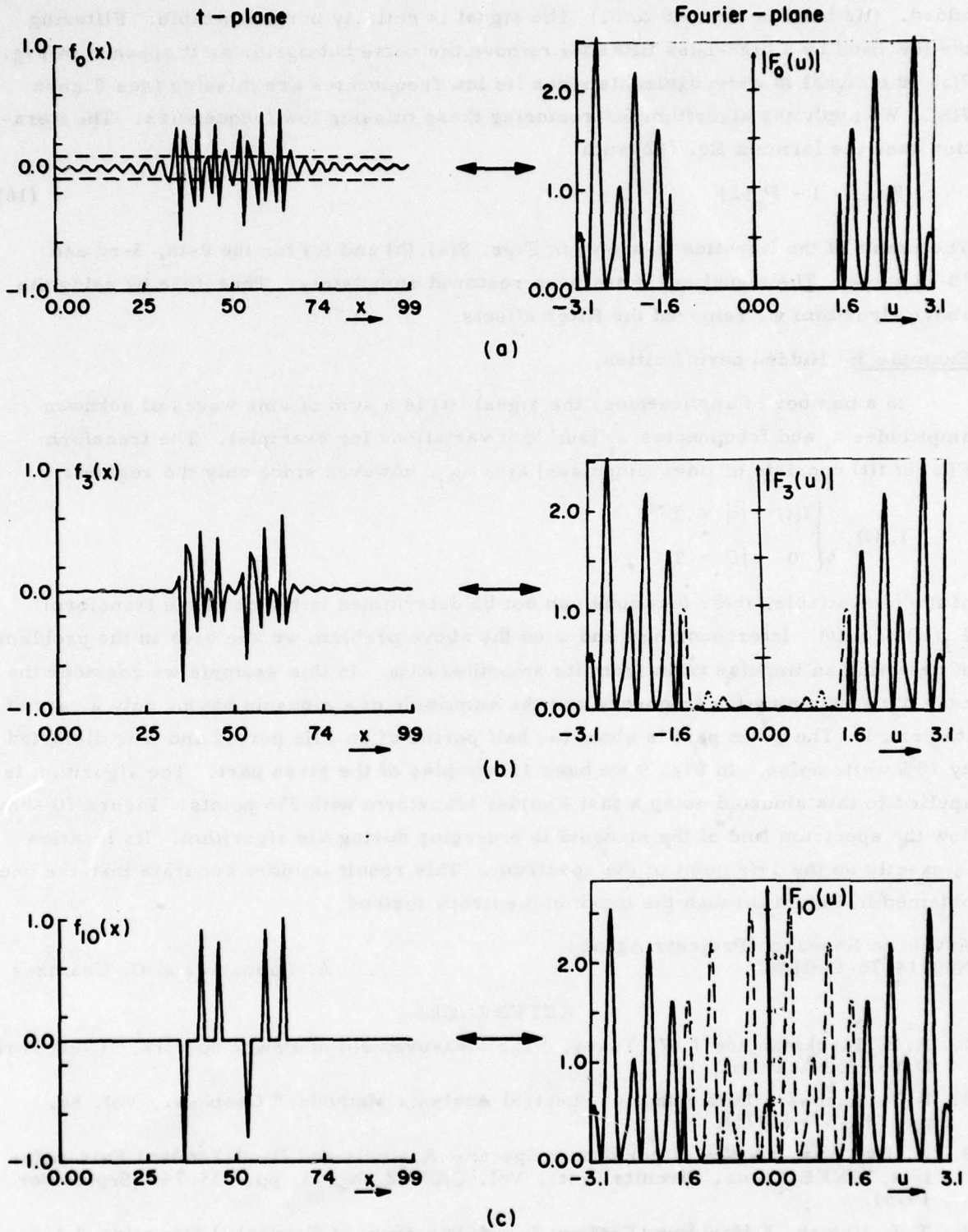


Fig. 8. The iteration for restoring the missing low frequencies.
(a) The signal $f_0(t)$ after the high pass filtering.
(b) The result of the 3-rd step of the iteration.
(c) The result of the 10-th step of the iteration.
(The signal has been restored completely.)

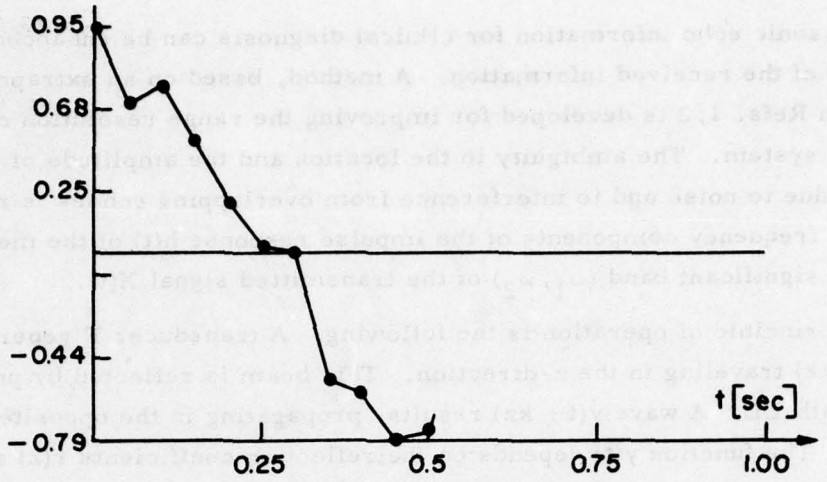


Fig. 9.

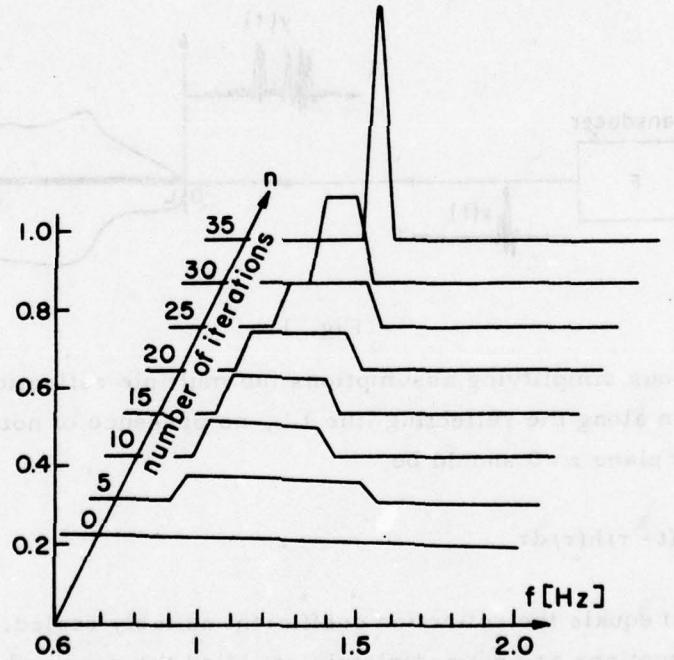


Fig. 10. The iteration.

HIGH RESOLUTION IN PULSE-ECHO SYSTEM

A. Papoulis and C. Chamzas

Ultrasonic echo information for clinical diagnosis can be enhanced by suitable processing of the received information. A method, based on an extrapolation scheme proposed in Refs. 1, 2 is developed for improving the range resolution of an ultrasonic pulse-echo system. The ambiguity in the location and the amplitude of the return signal y(t) due to noise and to interference from overlapping echoes is reduced by restoring the frequency components of the impulse response h(t) of the medium that are outside the significant band (ω_1, ω_2) of the transmitted signal X(t).

The principle of operation is the following. A transducer F generates an acoustic beam x(t - kz) traveling in the z-direction. This beam is reflected by parts of the body along its path LL. A wave y(t + kz) results, propagating in the opposite direction (see Figure 1). The function y(t) depends on the reflection coefficients r(z) along LL and is used to estimate r(z). The instruments in current use are based on simple schemes for estimating r(z) (envelope detection).

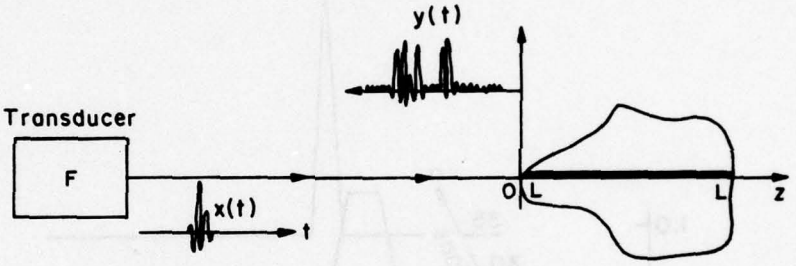


Fig. 1.

Under various simplifying assumptions (no multiple reflections, no attenuation, no beam variation along the relfecting line LL, no presence of noise, etc.) the reflected signal at the exit plane z = 0 should be

$$s(t) = \int_{-\infty}^{\infty} x(t-\tau)\,h(\tau)\,d\tau \tag{1}$$

where h(t) = r(kz) equals the reflection coefficient suitably scaled. However, since the simplifying assumptions are not completely satisfied the received signal y(t) is

$$y(t) = s(t) + \underline{n}(t) \tag{2}$$

where $\underline{n}(t)$ in the present work is assumed to be white noise uniformly distributed.

Taking Fourier transforms of both sides of Eq. (2) we obtain

$$Y(\omega) = X(\omega)H(\omega) + \underset{\sim}{N}(\omega) \tag{3}$$

Since x(t) is a modulated acoustic wave with a spectrum $X(\omega)$ negligible outside an interval $(\omega_1\ \omega_2)$ (see Fig. 2) the problem of estimating $H(\omega)$ can not be solved by simply deconvolving $Y(\omega)$ in Equation (3). Hence, due to the presence of noise, $H(\omega)$ can be reliably estimated only for ω inside this interval.

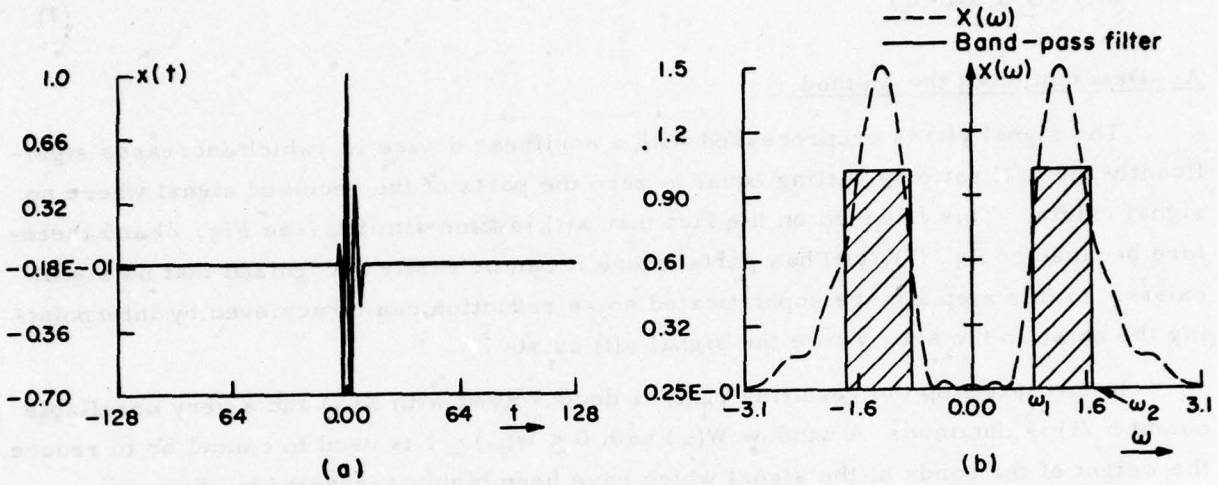


Fig. 2. The output x(t) of the transducer and its Fourier transform pair.

Let

$$H_a(\omega) = \frac{Y(\omega)}{X(\omega)} - \frac{\underset{\sim}{N}(\omega)}{X(\omega)} \tag{4}$$

be the estimator obtained by deconvolving Equation (3). To reduce the resulting error we must either replace $H_a(\omega)$ by zero outside the reliable interval (ω_1, ω_2) or multiply by a suitable window $W(\omega)$ that weighs properly the estimator $H_a(\omega)$. Thus as a first estimator we have (see Fig. 3)

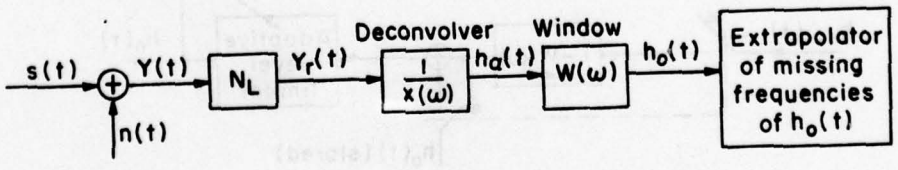


Fig. 3.

$$H_o(\omega) = H_a(\omega)\, W(\omega) \tag{5}$$

In the present work a method is described where the missing frequencies of $H_o(\omega)$ can be recovered if the reflecting substance consists of discontinuity surfaces. In this case

$$h(t) = \sum_i a_i \delta(t - t_i) \tag{6}$$

hence

$$s(t) = \sum_i a_i x(t - t_i) \tag{7}$$

A. Description of the Method

The signal y(t) is preprocessed with a nonlinear device N_L which increases significantly the S/N ratio by setting equal to zero the parts of the received signal where no signal exists. This is based on the fact that x(t) is time-limited (see Fig. 2) and therefore because of Eq. (7), y(t) has parts where it can be easily recognized that no signal exists. In this step a more sophisticated noise reduction can be achieved by interpolating the noise in the area where the signal s(t) exists.[3]

In the next step the resulting $y_T(t)$ is deconvolved with $X(\omega)$ and a very unreliable output $h_a(t)$ is obtained. A window $W(\omega)$ with $0 \leq W(\omega) \leq 1$ is used to cancel or to reduce the weight of the bands of the signal which have been highly corrupted by the noise. These bands, as we can see from Eq. (4), are the frequencies outside the (ω_1, ω_2) interval. Notice that by using a window of the form

$$W(\omega) = X(\omega)\, W_1(\omega) \tag{8}$$

the undesired deconvolution can be avoided (see Equations (4) and (5)).

The first estimator $h_o(t)$ of h(t) is fed in the extrapolator where the missing or reduced frequencies of h(t) are recovered. The extrapolator (see Fig. 4) performs an iteration by using the following scheme.

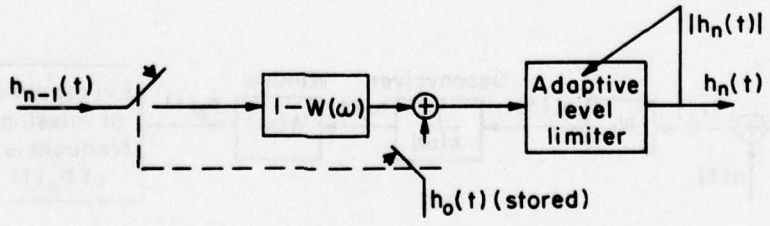


Fig. 4. The extrapolator.

It forms the function

$$H_{n-1}(\omega) = H_o(\omega) + [1 - W(\omega)]\, G_{n-1}(\omega) \quad 0 \leq W(\omega) \leq 1$$

and then sets equal to zero the parts of the signal which are below a certain level. That is

$$g_n(t) = \begin{cases} h_n(t) & \text{if} \quad |h_n(t)| > C_n \\ 0 & \text{if} \quad |h_n(t)| < 0 \end{cases} \tag{9}$$

In the beginning of the iteration C_n has a very low value C_o but during the processes C_n is increased as follows:

If

$$K_n = \min[\{|g_{n-1}(t)|\} - \{0\}]$$

then

$$C_n = \begin{cases} K_n & K_n < C_{max} \\ C_{max} & \text{if} \quad K_n > C_{max} \end{cases} \tag{10}$$

C_o and C_{max} are two constants which are calculated before the iteration starts. Noise level and expected value of minimum reflection coefficient a_i are the basic criteria for defining C_o and C_{max} correspondingly. Therefore utilizing the additional information that the reflecting substance consists of discontinuity surfaces, an iterative algorithm has been created which converges fast and moreover is quite stable in a noisy environment. The above iteration has successfully been applied in different cases. The computations are carried out digitally in terms of discrete Fourier series.

B. Illustration

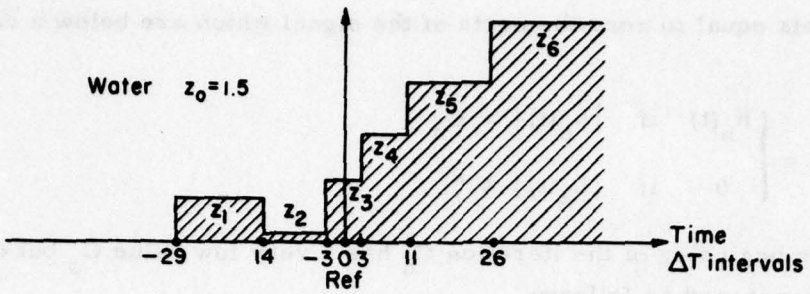
Consider a synthetic laminate comprised of six reflecting boundaries, related to the impedance discontinuities at each of these boundaries. We can use the values of these reflection coefficients to form a series of reflective echoes, i.e.,

$$s(t) = \sum_{i=1}^{6} a_i\, x(t - t_i) \tag{11}$$

where a_i = reflection coefficient occurring at boundary i at a time t_i. The synthetic laminate chosen for this illustration is shown in Figure 5. The resulted h(t) is shown in Figure 6. Figure 7 shows the synthesized echo train s(t) and Fig. 8 shows the received echo y(t) where

$$y(t) = s(t) + \underline{n}(t)$$

with $\underline{n}(t)$ white noise uniformly distributed in [-0.013, 0.013]. The signal $y_T(t)$ is



Reflection a_i	0.085	-0.08	0.1	0.08	0.09	0.1
Impedance $\frac{Z_i}{Z_o}$	1.186	0.852	1.222	1.174	1.199	1.222

Reflection coefficients $a_i = \frac{Z_i / Z_o - 1}{Z_i / Z_o + 1}$

Fig. 5.

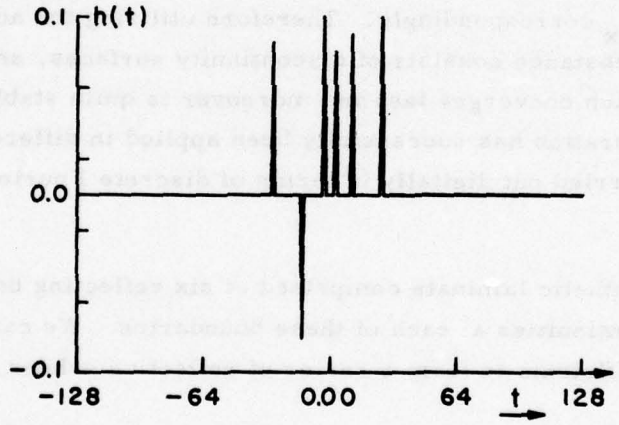


Fig. 6.

obtained by setting equal to zero the part of y(t) outside the lined area in Figure 8. The impulse response $h_a(t)$ resulting by deconvolution is shown in Figure 9.

A window is used as a band-pass filter (lined area in Fig. 2 or Figure 10(a)). The first estimate $h_o(t)$ is shown in Figure 10(b). The signal $h_o(t)$ is almost noise free ($S/N \cong 52$) but is highly distorted because of the filtering by $W(\omega)$. Figure 11(a) shows different steps of the iteration and Fig. 11(b) shows the result of the 40-th step.

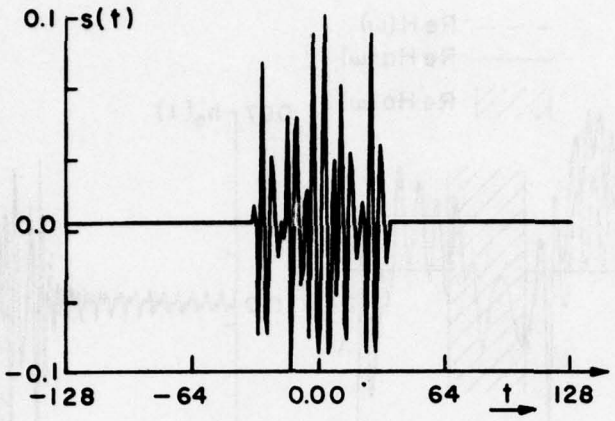


Fig. 7.

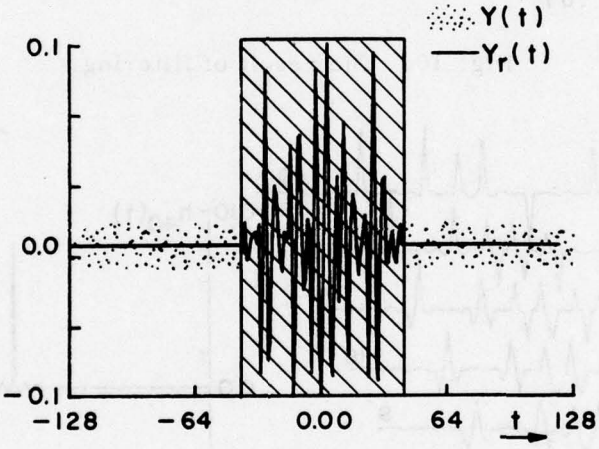


Fig. 8.

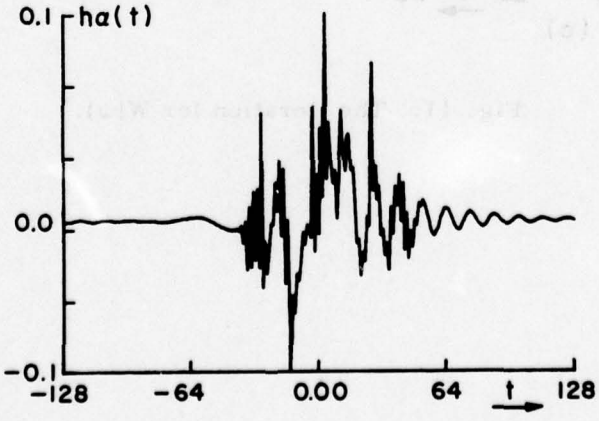


Fig. 9. The result of direct convolution.

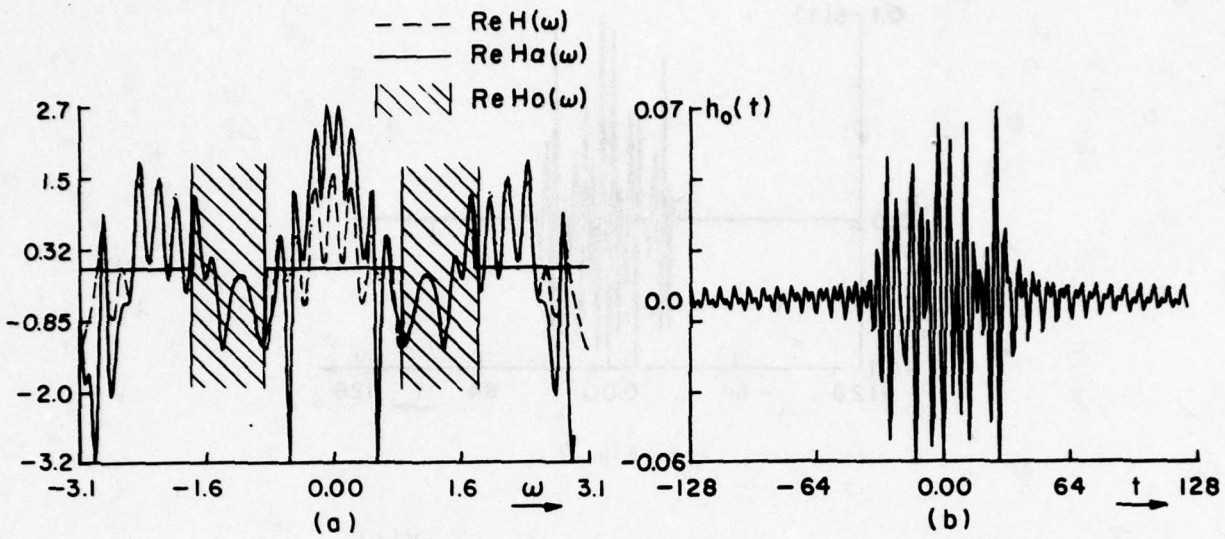


Fig. 10. The result of filtering.

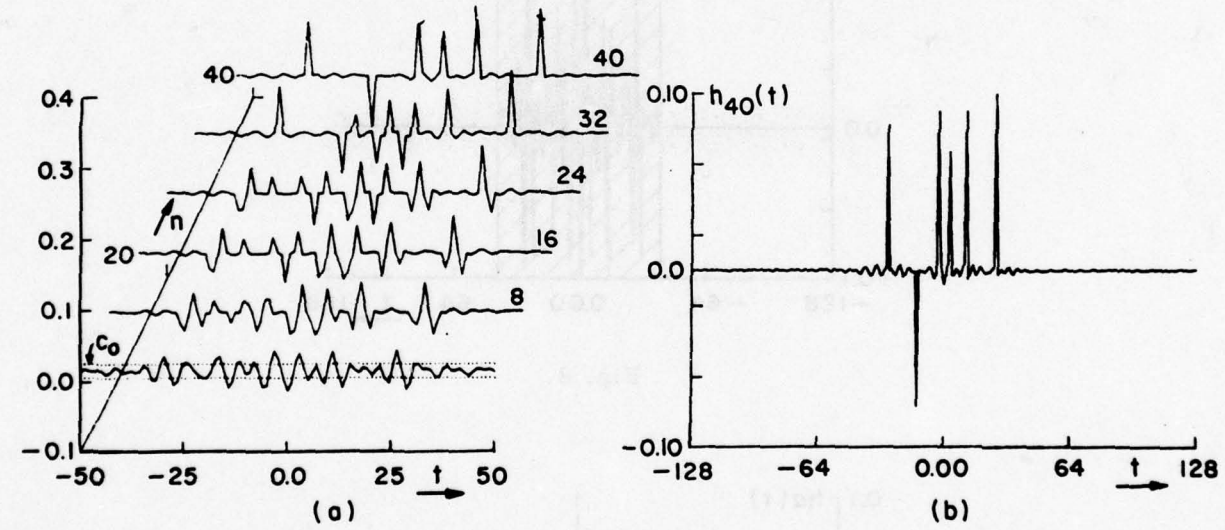


Fig. 11. The iteration for W(ω).

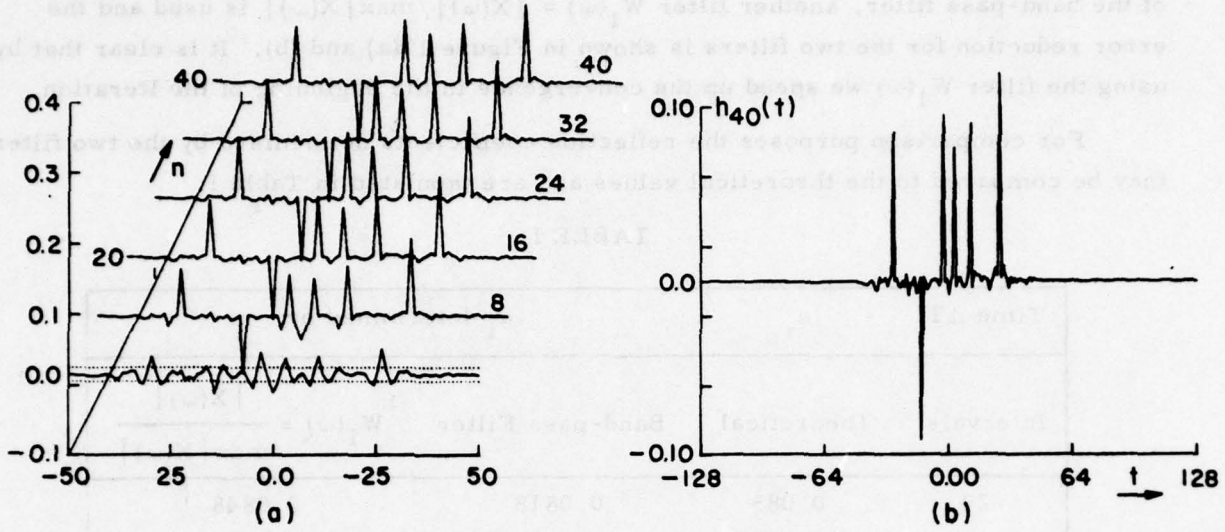


Fig. 12. The iteration for $W_1(\omega)$.

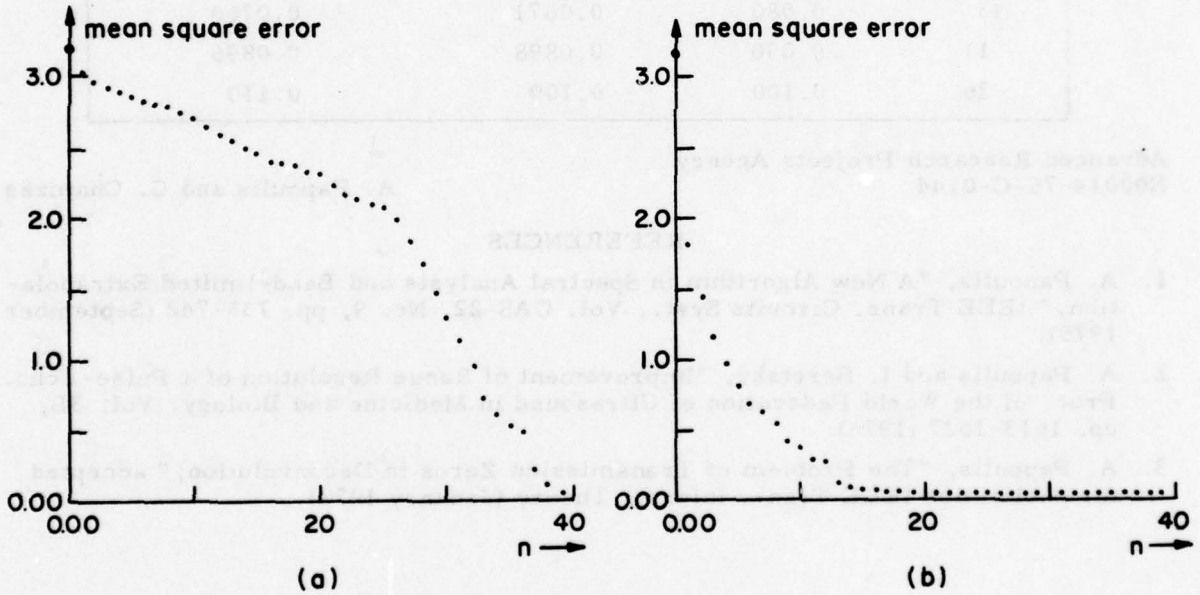


Fig. 13. Mean square error for the two different windows.

$$\text{(a)}\quad W(\omega) = \begin{cases} 1 & \text{for} \quad \omega_1 < |\omega| < \omega_2 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{(b)}\quad W_1(\omega) = \frac{|X(\omega)|}{\max |X(\omega)|}$$

Figure 12(a), (b) shows the same steps of the iteration and the $h_{40}(t)$ when, instead of the band-pass filter, another filter $W_1(\omega) = |X(\omega)|/\max|X(\omega)|$ is used and the error reduction for the two filters is shown in Figure 13(a) and (b). It is clear that by using the filter $W_1(\omega)$ we speed up the convergence in the beginning of the iteration.

For comparison purposes the reflection coefficients determined by the two filters may be compared to the theoretical values and are tabulated in Table I.

TABLE I.

Time ΔT	a_i	a_i determined by	
Intervals	Theoretical	Band-pass Filter	$W_1(\omega) = \dfrac{\lvert X(\omega)\rvert}{\max\lvert X(\omega)\rvert}$
-29	0.085	0.0818	0.0848
-14	-0.080	-0.0754	-0.0903
-3	0.100	0.0893	0.0952
+3	0.080	0.0671	0.0760
11	0.090	0.0898	0.0896
26	0.100	0.100	0.110

Advanced Research Projects Agency
N00014-76-C-0144

A. Papoulis and C. Chamzas

REFERENCES

1. A. Papoulis, "A New Algorithm in Spectral Analysis and Band-limited Extrapolation," IEEE Trans. Circuits Syst., Vol. CAS-22, No. 9, pp. 735-742 (September 1975).

2. A. Papoulis and I. Beretsky, "Improvement of Range Resolution of a Pulse-Echo," Proc. of the World Federation of Ultrasound in Medicine and Biology, Vol. 3B, pp. 1613-1627 (1976).

3. A. Papoulis, "The Problem of Transmission Zeros in Deconvolution," accepted for publication IEEE Trans. Inform. Theory (January 1978).

A ROBUSTIZED KALMAN FILTER

I. Kadar and L. Kurz

Recently there has been considerable interest in recursive estimators for the linear model that protect the estimate from outliers and are robust. These include the use of Huber's M-estimators[1] light-limiter robustized stochastic approximation (SA) estimates based upon Huber's results,[2] SA estimators both of the Robbins-Monro and Kiefer-Wolfowitz type with adaptive gain coefficients for scalar case and adaptive gain matrices for the vector case,[3,4] with applications to robust SA estimates using non-parametric statistics[3] and general (not necessarily robust) estimation theory based upon discrete-time point processes using martingales.[6] The desired robust quality of the estimator is usually assessed in terms of not only the extent to which it removes outliers in the presence of contamination, but rather, by its performance in unknown, possibly asymmetric, noise environment[7] reaching asymptotic normality with practically moderate sample sizes, unbiased and perhaps optimal (although not necessarily BLUE) in some sense when constructed for one kind of noise environment, and in fact, another noise environment holds.

Most of the above cited work on robust estimation for the linear model has been applied to either point estimation (e.g., Refs. 1 and 7) or to detection, Ref. 3, 4, 9, with the exception of Ref. 8, where robustized one-step Bayesian like estimators were sequentially applied to robustize the Kalman Filter (KF) for either (symmetric heavy-tailed) non-Gaussian state or (symmetric heavy-tailed) non-Gaussian measurement noise densities. The work in Ref. 8 considered the application of M-estimates and min-max theory for the ϵ-contaminated normal family combined with results from influence function robustized SA-theory[2] using the linear-non-linear (L-N) robustizing approach.[10] The L-N approach, which applies first a linear transformation to scale and symmetrize the CDF of the innovations and then operates on the result with a non-linear (i.e., light-limiter) robustizing transformations, suffers from a major drawback. When the underlying noise CDF is unknown, it is impossible to find the linear transformation, so the CDF is assumed to be Gaussian, the linear transformation computed and the nonlinearity is applied with the hope that it will robustize the estimate when the CDF is non-Gaussian. Clearly, when the CDF is non-Gaussian, the assumed linear transformation is inappropriate and a number of approximations are needed in order to justify the results of Reference 8.

The material presented here is a generalization of the work of Masreliez and Martin[8] in robustizing the KF. We first establish the background for the L-N approach from MLE and KF point estimation viewpoint and show the asymptotic relationships between KF and SA. This leads directly to the one-step robust L-N estimator in

Theorem 1. We then show that it is possible to avoid the ambiguity in selecting the linear transformation by the use of a batch-integer-rank-linear preprocessing transformation on the innovations (batch-non-linear-linear B-N-L approach)[3] using the two-sample Mann-Whitney-Wilcoxon nonparametric statistics (MWWNS) from the Chernoff-Savage class[11] in batch processing (in the spirit and form of Evans, Kersten and Kurz[3] and Kersten and Kurz[4] generalizing Theorem 3 in Reference 8. We consider the case when the measurement noise probability density function (pdf) is described by the mixture distribution model

$$= \{f(\cdot);\ f(\cdot) = (1-\varepsilon)\,g(x) + \varepsilon\,h(x),\ h(\cdot)\varepsilon\,\mathcal{H},\ 0 \leq \varepsilon \leq 1\}$$

where g(x) is a given density function and $\mathcal{H}$ is a wide class of density functions, where $h(\cdot)\varepsilon\,\mathcal{H}$ is usually a high variance pdf representing a burst noise component of the measurement channel. The case of non-Gaussian state noise pdf (Theorem 4 in Ref. 8) is not considered since our generalized results in Theorem 2 (the B-N-L approach) directly apply. The proof of Theorem 2 is similar to methods in Refs. 3, 4 and 8, and follows from a significant result from the Chernoff-Savage theory on asymptotic normality and symmetry properties for a class of linear two-sample statistics[11] and its extensions,[12, 13] along with results for the particular case of the MWWNS which reaches asymptotic normality practically with as few as eight batch samples[14, 15] used in robust SA estimation of location.[5] This means that after a preprocessing delay of m-steps, the MWWNS robustized filter based upon the nominal Gaussian CDF's becomes asymptotically ($m \geq 8$) Gaussian and is essentially unaffected by the degree and nature of the contamination, i.e., $h(\cdot)\varepsilon\mathcal{H}$, by the limiting effect of the integer-rank transformation. So, in effect, we immunize the measurements against burst noise and our filter is robust for non-Gaussian noise.

In order to establish a relative comparison measure between L-N and B-N-L approaches, we convert the robust KF problem to a robust point estimation problem and apply asymptotic results from robust Robbins-Monro (RMSA) procedures based upon the work of Evans, Kersten and Kurz[2, 3] to compute variance bounds of the resultant estimates. This allows the computation of a criteria based upon the relative sample number ratio (RSNR) of the B-N-L and L-N approaches.

It should be noted that the vector KF is addressed directly in this report and we refer to the description of the general linear discrete-time model in Ref. 16 for the familiar minimum variance KF. Our direct solution of the vector case is also based upon the adaptation of some recent results on robust vector RMSA by Kersten and Kurz.[4]

Finally, Monte Carlo simulation results and numerical examples substantiate the theory by verifying the improved performance of the B-N-L filter robustized via the MWWNS over the L-N light-limiter robustized filter.

A. Robust Filtering Via the L-N Approach: Some Known Results

We intend to robustize the well-known linear model

$$y_k = H_k x_k + v_k \tag{1}$$

where y_k is a $(v \times 1)$ measurement vector, H_k is a $(r \times n)$ measurement matrix, x_k an $(n \times 1)$ state vector and v_k $(r \times 1)$ measurement noise. Furthermore, assume that the state, x_k evolves from

$$x_k = \phi_{k,k-1} x_{k-1} + w_{k-1} \tag{2}$$

where $\phi_{k,k-1}$ is $(n \times n)$ state transition matrix and w_k is the plant noise. With $\{v_k\}$ and $\{w_k\}$ zero mean mutually independent white-noise sequences which are also independent of the initial state x_o (which is also Gaussian) leads to the minimum variance Kalman Filter (KF) described in detail in Reference 16. As discussed in Ref. 16, the estimate, $\hat{x}_k$ at stage n is formed by the prediction $\bar{x}_k = \phi_{k,k-1} \hat{x}_{k-1}$ corrected by a term proportional to the innovation, $y_k - H_k \bar{x}_k$.

Now consider the case where the measurement noise is heavy-tailed and the plant noise is Gaussian. The large innovation vector in this case is due to the outliers in the measurement noise density. So, intuitively, we would want to attenuate the large innovations by some operation and place more emphasis on our predictions. This intuitive reasoning leads one directly to the robust one-step Bayesian-like estimator and multi-step application thereof, comprising the basis of the "L-N approach." In the following, we present a brief review leading to the development of the one-step Bayesian L-N estimator for the case of the ε-contaminated normal family to allow a direct means of comparison with the B-N-L approach realized in Theorem 2 in the next section. Our approach in this area is similar to Ho,[17] leading directly to Theorem 1.

1. Point Estimation (MLE and KF) and Stochastic Approximation

Consider the maximum likelihood estimate of x, given in Eqs. (1) and (2) with $H_k = H, \phi_{k,k-1} = [I]$ for all k, and $W_k = 0, \forall k$. It is well known that[17]

$$\hat{x}_k = \hat{x}_{k-1} + P_k H^T (y_k - H \hat{x}_{k-1}) \tag{3}$$

where $\hat{x}_o = 0$, P_o is a given $(n \times n)$ positive definite matrix,

$$E[v_K v_J^T] = I \delta_{kj}$$

and

$$P_k^{-1} = P_{k-1}^{-1} + H^T H \tag{4}$$

Applying the Sherman-Morrison Matrix Inversion Lemma[16] to Eq. (4), we obtain the KF equations as in Ref. 16 and applying the resultant Kalman gain term, i.e., $P_{k-1}H^T(HP_{k-1}H^T + I)^{-1}$ to itself, we obtain, following Ho[17] in the limit

$$\lim_{k\to\infty} P_k H^T = (1/k)\, P_o H^T (HP_o H^T)^{-1} \tag{5}$$

and we obtain asymptotically

$$\hat{x}_k = \hat{x}_{k-1} + (1/k)\, P_o H^T (HP_o H^T)^{-1} (y_k - H\hat{x}_{k-1}) \tag{6}$$

which is in the form of a multidimensional SA.[4] Now with SA and KF asymptotically related (for the point estimation case) it is quite natural to use asymptotic results from SA theory as bounds for the approximate KF structures.

The above analysis is directly applicable to robust SA using a general class of nonlinear regression functions (see recent results of Kersten and Kurz[4]). Furthermore, the previous asymptotic scalar result of Martin[2] for the soft-limiter robustized SA (Huber's M-estimator) case can now be extended rigorously to the vector case (see Kersten and Kurz[4]).

2. The One-Step Robust Bayesian Form

Now let in Eq. (3), $k = 1$, $\hat{x}_o = x$ and $P_o = M$, and introduce an odd symmetric nonlinearity, $\psi(\cdot)$, with

$$\psi(\cdot) = \ell_k(t) = \begin{cases} k, & t > k \\ t, & |t| \le k \\ -k, & t < -k \end{cases}$$

the light-limiter influence function operating on the innovation-like term, i.e., $\psi(y - H\bar{x})$ with the following two properties of the distribution of the innovation, $\nu = y - H\bar{x}$ assumed to hold:

(P1) $F_\nu(\nu_1, \nu_2, \cdots, \nu_i, \cdots, \nu_r) = 1 - F_\nu(\nu_1, \nu_2, \cdots, -(\nu_i + 0), \cdots, -\nu_r)$

(P2) All marginal distributions for F_ν have absolutely continuous densities and are members of F with $F \in F_\varepsilon$ where

$$F_\varepsilon = \{F(\cdot)/F = (1-\varepsilon)N + \varepsilon H,\ 0 \le \varepsilon \le 1,\ H \text{ symmetric}\}$$

and N is the unit normal CDF, and we obtain the Bayesian one-step robust estimator:

Theorem 1:[8] Consider the measurement relation, Eq. (1), and assume that the CDF, F_ν of $\nu = y - H\bar{x}$, satisfies (P1) and (P2). Let the prior density for x, $\bar{x}$ be Gaussian $n(\cdot/\bar{x}, M)$. Then the estimator

$$\hat{x} = \bar{x} + MH^T \psi(\nu) \tag{7}$$

will yield a bounded variance

$$P = E[(x - x)(\hat{x} - x)^T] \leq M - MHH^T M E_{F_o}[\psi_F'] \tag{8}$$

where $\psi(\cdot)$ is an (r x 1) vector operator with components $\psi_i(t) = \psi_F(t_i)$ where $\psi_F(\cdot)$ is the influence function corresponding to the univariate family F. $E_{F_o}[\cdot]$ denotes expectation for the least favorable pdf in F and $\psi_F' \equiv d\psi_F/dt$.

Proof: The proof is given in Ref. 8 and in part of the proof of the robust filter realized in Theorem 2 in the next section. It follows by direct substitution and by the application of a bound on the asymptotic variance of the light-limiter robustized SA estimator coupled with Huber's min-max result for the ε-contaminated normal family with one added crucial assumption not mentioned in Reference 8. For the vector case under consideration, the robust SA vector estimator used in establishing the asymptotic bound has to be decoupled with a diagonal gain matrix and the light-limiter nonlinearity applied component-wise to the regression function. So in essence, we have r SA procedures operating in parallel. Then Huber's results adapted by Martin[2] follow component-wise, i.e., with a diagonal covariance matrix whose elements are

$$V_{SA_i}(T, F) = \frac{A_i^2 E_F(\psi_i^2)}{2 A_i E_F(\psi_i') - 1}$$

where $A_i = F_o^{-1}(\psi_i')$ and $\psi_i = -f_{oi}'/f_{oi}$ with f_o the least favorable pdf in F which leads to a SA estimator, T_o, such that $V_{SA_i}(F, T_o) \leq V_{SA_i}(F_o, T_o)\ \forall$ i for $F \varepsilon F_\varepsilon$ where $V_{SA_i}(F_o, T_o) = E_{F_o}^{-1}(\psi_i')$.

The statement of Theorem 1 and its proof[8] assumed that the CDF of the innovation vector satisfied the symmetry and scale conditions described in properties (P1) and (P2). Note, however, that this is seldom the case and one has to first find a linear transformation, T, on the innovation vector, i.e., $\nu = T(y - H\bar{x})$ and then apply the nonlinearity (the L-N approach) having satisfied properties (P1) and (P2). This requires a priori knowledge of the CDF of the innovation vector which is somewhat contrary to the notion of robustness. The situation becomes even more complicated (in Theorem 3 of Ref. 8) where the one-step Bayesian estimator is sequentially applied to form a robust approximate filter which provides protection against measurement noise outliers. Now, T becomes T_K and it has to be recomputed at each step. This is the basic disadvantage of the "L-N approach," as pointed out before. However, the above disadvantage can be eliminated by replacing the L-N operation by an m-step B-N-L preprocessor

using the symmetric version of the MWWNS in batch processing. We avoid the ambiguity of selecting T_K and we asymptotically (after m-preprocessing steps) approximate a robust Gaussian filter resulting in a true robust structure whose operation is essentially independent of the underlying CDF's. This structure is realized in Theorem 2 of the next section.

B. The Batch Nonlinear-Linear (B-N-L) Preprocessing Approach

Consider the sequential application of the basic one-step Bayesian estimator delineated in Theorem 1 with the L-N operation changed to the B-N-L preprocessor operation. This consists of forming a batch of samples of size m (of the components of the innovation vector) and selecting a nonparametric statistic of the form:

$$W^{m_1,m_2}(\cdot) = (1/m_1 m_2) \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \operatorname{sgn}(X_i - Y_j)$$

which is the symmetric version of the two sample Mann-Whitney-Wilcoxon nonparametric statistics (MWWNS) from the Chernoff-Savage class. However, before we state the B-N-L approach realized in Theorem 2, it is necessary to briefly examine some important characteristics of the MWWNS. One of the important and useful characteristics of the MWWNS is that its use is not restricted to any symmetry properties on the CDF's of the i.i.d. samples of

$$X_1, X_2, \cdots, X_{m_1} \quad \text{and} \quad Y_1, Y_2, \cdots, Y_{m_2}$$

with continuous CDF's F(x) and G(x) respectively, with a number of additional useful properties (see Appendix A of Ref. 18) under, H, (F(x) = G(x))

$$E[W^{m_1,m_2}] = 0$$

$$\operatorname{Var}[W^{m_1,m_2}] = (m_1 + m_2 + 1)/3 m_1 m_2$$

Also, under H and K ($F(x) \neq G(x)$)

$$\sup_{F,G} W^{m_1,m_2} = -\inf_{F,G} W^{m_1,m_2} = 1$$

With $W^{m_1,m_2}(\cdot)$ a member of the Savage-Chernoff class of two-sample linear rank statistics, it can be shown to be equivalent to a general two-sample statistic of the form

$$T_N(Z) = \sum_{i=1}^{N} a_{Ni} Z_{Ni}$$

(see Appendix A of Ref. 18) with $N = m_1 + m_2$, $Z = (Z_1, Z_2, \cdots, Z_N)$ a set of indicator variables, where $Z_i = 1$ if the ith random variable from the combined ordered sample of the X's and the Y's is an X and $Z_i = 0$, otherwise for $i = 1, 2, \cdots, N$, and with scores $a_{Ni} = i$. Now the asymptotic theory of Chernoff and Savage[11] and its extensions[12,13] apply to W^{m_1, m_2} under both $F(x) = G(x)$ and $F(x) \neq G(x)$,

$$\lim_{N \to \infty} P\left[\frac{W^{m_1, m_2} - EW^{m_1, m_2}}{\sqrt{\text{Var } W^{m_1, m_2}}} \leq t\right] = \Phi(t)$$

where $\Phi(t)$ is the unit normal CDF. The above result along with symmetry properties of CDF of $T_N(Z)$ makes this statistic ideally suited for the B-N-L approach. Furthermore, from a practical point of view, it is shown by Mann and Whitney[14] and Fix and Hodges[15] that asymptotic normality is achieved to an accuracy within three places with $m_1 = m_2 = m = 8$. This means that after a preprocessing delay of m-steps (with $m \geq 8$) we have reached robustness, as mentioned before. In addition, the MWWNS is the most effective as measured by its Asymptotic Relative Efficiency (ARE) under H when the distribution has heavy tails, i.e., impulse noise. With T_1 and T_2 two different non-parametric test statistics with n_1 and n_2 as dimensions

$$\text{ARE}_{T_1/T_2} \equiv \lim_{n_1, n_2 \to \infty} (n_2/n_1)$$

with the significance level and power of the test fixed (compare ARE to RSNR in Section C). Intuitively, this property can be attributed to the sign $(\cdot)$ function. So, in essence, the robust B-N-L filter is also a robust non-Gaussian filter due to the m-sample asymptotic properties of the MWWNS. It should also be noted that other two-sample rank statistics from the Chernoff-Savage family could also be used in place of the MWWNS; however, the MWWNS is simple to implement and has good (moderate sample) asymptotic properties which are well established.

1. Implementation of the B-N-L

From a practical implementation point of view, B-N-L preprocessing operation requires $k = m$ initial iterations of the robust filter to obtain the m-batch samples. This is accomplished by initially storing the null vector of length m in the batch variables and replacing the elements of the null vector by shifting the data into the batch variables at each iteration step. Therefore, while the batch is "filling up," the method requires that we initially assume that the distribution of the prediction density and the nominal measurement noise density is Gaussian, which becomes justified after the preprocessing delay of m-steps, due to the m-sample asymptotic normality stated before. This

assumption is identical to the one used in the L-N approach, albeit the required assumptions needed to implement the L-N scheme are never completely justified since the limiter operation merely removes outliers and does not possess the required asymptotic properties.

Now we are in a position to present the following theorem for the case when we desire protection against measurement noise outliers.

Theorem 2: Consider the linear model, Eqs. (1) and (2) and make the following assumptions:

(A1) $(x_k - \bar{x}_k) \equiv \tilde{x}_k$ is Gaussian with mean zero and covariance matrix, M_k $\forall$ k.

(A2) A sequence of diagonal scaling matrices is known such that

$$S_k = A_k \begin{bmatrix} W^m_{k_1}(y_{k_1} - \xi_{k_1}) \\ W^m_{k_2}(y_{k_2} - \xi_{k_2}) \\ \vdots \\ W^m_{k_r}(y_{k_r} - \xi_{k_r}) \end{bmatrix} \tag{9}$$

where $\xi_k = H_k \bar{x}_k$, satisfies properties (P1) and (P2) stated before, $W^m_k(\cdot)$ is a (r x 1) vector m-batch-integer-rank operator with scalar components using a symmetric version of the Mann-Whitney-Wilcoxon statistics*[3]

$$W^m_{k_s}(\cdot) = (1/m^2) \sum_{i=1}^{m} \sum_{j=1}^{m} \operatorname{sgn}\{y_{[i+m(k-1)]_s} - \xi_{[j+m(k-1)]_s}\} \tag{10}$$

and

$$A_k^{-1} = \begin{bmatrix} \sigma_1 & & & 0 \\ & \sigma_2 & & \\ & & \ddots & \\ 0 & & & \sigma_r \end{bmatrix} \tag{11}$$

where

$$\sigma_1 = \sigma_2 = \cdots = \sigma_r = \sigma = \sqrt{(2/3) + 1/3m^2}$$

*W of Eq. (10) and A of Eq. (11) are independent of k. In more complicated problems they may be functions of k.

Then the filter

$$\hat{x}_k = \bar{x}_k + M_k H_k^T S_k$$
$$\bar{x}_k = \phi_{k,k-1}\hat{x}_{k-1} \tag{12}$$

yields an unbiased estimate of $\hat{x}_k$ of the state x_k with M_k defined by

$$M_k = \phi_{k,k-1} P^*_{k-1} \phi^T_{k,k-1} + Q_{k-1} \tag{13}$$

$$P^*_k = M_k - M_k A_k E_{F_o}[W'_k(\cdot)]\, A_k^T H_k M_k \tag{14}$$

where $E_{F_o}[W'_k(\cdot)]$ is a diagonal matrix with elements

$$\alpha_{k_i} = f_{v_{k_i} - \tilde{\xi}_{k_i}}(0), \quad i = 1, 2, \cdots, r$$

where $f(\cdot)$ is the one-dimensional nominal Gaussian density of the components of $v_k - H_k \tilde{x}_k$, will yield a bounded error covariance matrix

$$P_k = E[(\hat{x}_k - x_k)(\hat{x}_k - x_k)^T] \leq P^*_k \tag{15}$$

for all k.

Proof: See Appendix A of Reference 18.

2. Generalized Gaussian Noise Approximation

It follows from the statement of Theorem 2 that the worst case density no longer exists when we completely robustize the innovation vector, so the scalar terms $\alpha_1 = f_{vi - \tilde{\xi}_i}(0)$ can be precomputed for $i = 1, 2, \cdots, r$. As a matter of fact, a simple approximation can be derived in precomputing the α_i's, such that $\alpha_i \simeq \alpha_{ki} = \alpha = f(0)$ where f(0) is the nominal Gaussian pdf of the innovation.

Assume that the pdf of each component of the innovation vector v_{ki} is given by

$$g_c(v) = \frac{c}{2A(c)\Gamma(1/c)} \exp\{-[|v|/A(c)]^c\}$$

where

$$A(c) = [\sigma^2 \Gamma(1/c)\Gamma(3/c)]^{1/2}$$

$\Gamma(\cdot)$ is the gamma function. The variance of $g_c(v)$ is σ^2 and $g_c(\cdot)$ is a generalization of the Gaussian density function with zero mean and variance σ^2, the resulting class of

densities being symmetric and unimodal with different rates of exponential decay. (For $c = 2$, $g_c(\cdot)$ is the Gaussian density; for $c = 1$, $g_c(\cdot)$ is the Laplace density and for other values of c, $g_c(\cdot)$ approximates a large class of both thin and heavy tailed densities.) Now in examining the variation of $g_c(0)$ with c, we find that with $1 \leq c \leq 4$, $g_c(0)$ varies only in the range of 2 to 1. So one would expect the MWWNS robustized filter to be insensitive to the exact value of f(0). As a matter of fact, simulation results (in mixture noise with 10% burst noise component of variance 64) indicate that with a 200% change in f(0) the state variance of the robust filter only changed 0.1% (from 0.12770 to 0.12634). This result further confirms the robust property and simplifies the construction of the B-N-L filter for the contaminated Gaussian measurement noise family $\exists[\varepsilon, R_k]$.

3. Properties of the B-N-L in non-Gaussian and Asymmetric Contaminated Noise

An important result which follows directly from examining the properties of the MWWNS is that the contaminated measurement noise need not be Gaussian (as stated in Section A) as long as both the given density function and the contamination are symmetrically distributed. For the case of asymmetric contamination, the MWWNS becomes only median unbiased, as shown by Hodges and Lehmann[19] and, $\hat{x}_k$ will no longer be an unbiased estimate of the state x_k. In this case one could increase the dimensionality of the state vector by augmenting the state by a constant, i.e., $b_k = b_{k-1}$ and measurement equations by $y_k = H_k x_k + v_k + b_k$, and estimate the bias, b, along with the other parameters.

$$\begin{bmatrix} x_k \\ b_k \end{bmatrix} = \begin{bmatrix} \phi_{k,k-1} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x_{k-1} \\ b_{k-1} \end{bmatrix} + \begin{bmatrix} I \\ 0 \end{bmatrix} W_{k-1}$$

$$y_k = [H_k \quad I] \begin{bmatrix} x_k \\ b_k \end{bmatrix} + V_k$$

Therefore, we use the estimated state to compensate for the unknown bias.

C. A Relative Comparison Measure

In comparing the L-N and the B-N-L robust structures, it would be desirable to establish a relative comparison measure based upon the relative sample number ratios (RSNR) of the respective approaches and possibly compare each approach with the nominal minimum variance KF. This requires the computation of bounds on the asymptotic variances. In the case of the KF structure, this is only possible for the case of constant coefficients (the point estimation case) when the solution of a Ricatti equation governing the state covariance becomes a quadratic[16] and the exact asymptotic bound can be precomputed. However, the solution is still very difficult, except for simple

first order systems. Since we are only interested in a relative comparison measure, we take an alternate approach.

As illustrated in Section A, we evolved the robust approximate KF structures via the robust RMSA approach. So it is quite natural to refer to the asymptotically equivalent robust RMSA when we examine asymptotic behavior. Evans[5] has shown for the case of nonparametric statistics robustized scalar RMSA with adaptive gain coefficients, that the asymptotic variance of the resultant estimator satisfies

$$V_{SA}(T,F) = \frac{Var_F(\psi)}{n\,E_F^2(\psi')}\left[\frac{A^2 E_F^2(\psi')}{2A\,E_F(\psi') - 1}\right] \geq \frac{Var_F(\psi)}{n\,E_F^2(\psi')}$$

where ψ stands for the nonlinearity (nonparametric statistic) satisfying the constraints of Theorem 1 of Reference 3. For the vector RMSA with adaptive gain coefficient the gain coefficient becomes a diagonal gain matrix and the nonlinearity is applied componentwise as shown by Kersten and Kurz.[4]

Now, by equating the lower bounds with the explicit assumption that the covariance matrix of the resultant asymptotically normal pdf is diagonal

$$n_\ell\, V_{SA_{ii}}(T_\ell, F) = n_w\, V_{SA_{ii}}(T_w, F)$$

componentwise where, T_ℓ is the limiter influence function robustized RMSA given by Martin[2] and T_w stands for the MWWNS robustized RMSA given in Ref. 3, we generalize the results of Refs. 2 and 3 to the vector case and can solve for the RSNR with elements $\{V_{SA_{ii}}\}$. In the general case, an alternate measure can be defined in terms of the sum of the diagonal elements, i.e.,

$$n_\ell\, tr[V_{SA}]_{T_\ell} = n_w\, tr[V_{SA}]_{T_w}$$

which can be used for the general covariance matrix relations given in Ref. 4 by neglecting the contribution of the off-diagonal elements in determining RSNR.

Now under the simplifying assumptions of a diagonal covariance matrix, for the vector case, we obtain for the robust KF structures considered (for symmetric pdf's)

$$RSNR(T_\ell/T_w) = \frac{4f^2(0)\int_{-\infty}^{\infty} \ell_k^2 f(z)\,dz}{k^2\left[\frac{2}{3} + \frac{1}{3m^2}\right]\left[\int_{-k}^{k} f(z)\,dz\right]^2} \tag{16}$$

where $f(\cdot)$ is the pdf of the components of the innovation vector, ℓ_k is the light limiter influence function with scale parameter, k, and m is the batch size.

In the scalar case with Gaussian f(·) with unity variance, $k \ll 1$, and $H_k = I \Psi k$ Eq. (16) can be approximated as

$$RSNR(T_f/T_w) \simeq 12/(4\pi)^{3/2}$$

This result allows the comparison of the filters in terms of the computation time which is directly related to RSNR. Comparing the performance of the MWWNS for two different batch sizes, $m_1 = 10$ and $m_2 = 20$

$$RSNR(T_w^{m_1}/T_w^{m_2}) \simeq 1 \quad \text{for} \quad m > 10$$

which shows that we reach the asymptotic values rather rapidly independent of m, $m \leq 10$.

D. Simulation Results

A series of Monte Carlo simulations and numerical computations were performed to compare the performance of the L-N and B-N-L robustized filters with the KF in both Gaussian noise and in mixture noise. Both the robust (L-N and B-N-L) filters and the KF were designed assuming Gaussian CDF's with constant variances for the measurements, the plant noise and the initial state estimate, $v_k \sim N(0,1)$, $w_k \sim N(0,\sigma_q)$ and $x_0 \sim N(0,M_0)$, respectively. This allowed the comparison of performance of the filters when the simulated "unknown" measurement noise was different from Gaussian.

We used the first-order system model given in Ref. 8

$$x_k = 0.5\, x_{k-1} + w_{k-1}$$

$$y_k = x_k + v_k$$

with the density of v_k being the Gaussian mixture

$$f(v) = CN(\cdot/\varepsilon,\sigma) \equiv (1-\varepsilon)\, N(0,1) + \varepsilon\, N(0,\sigma),\ 0 \leq \varepsilon \leq 1,\ \sigma > 1$$

to facilitate direct verification of the results for the L-N case and to serve as a basis for comparison with the B-N-L approach. An advantage of using the above scalar model is that in this case (with constant coefficients) the asymptotic variances can be precomputed (because the solution of the Ricatti equation becomes a simple quadratic). This allows direct comparison of the Monte Carlo results with KF theory.

The L-N filter was implemented (using Theorem 3 of Ref. 8), where we used a light-limiter gain coefficient of k = 0.317 which locates k at one standard deviation of the Gaussian innovation CDF. In order to normalize the comparison between the L-N and the B-N-L approaches, we limited the excursion of MWWNS to $k = \pm 0.317$. For the B-N-L approach, we used batch sizes of m = 10 and m = 20 to verify the asymptotic

theory of MWWNS. We investigated the performance of the robust filters and the KF for different ratios of the plant noise and measurement noise variances

$$r \equiv \frac{\sigma_q^2}{(1-\epsilon) + \epsilon\,\sigma^2}$$

with $\epsilon = 0.1$ and $\sigma = 8$, representing a mixture noise with a 10% burst noise component with variance 64. The range of input parameter values used is listed in Table I. A Monte Carlo step size of 1000 iterations was used in all cases. We verified that the filters reached steady state by varying the value of the initial state covariance, M_0, between 0.1 to 1 and found no discernible change in the state variances within six significant figures.

TABLE I. Simulation input parameters.

Case	R	F(v)	σ_q^2	σ^2
1	0.01	CN	0.073	64
2	1	CN	7.3	64
3	0.01	N	0.073	0
4	1	N	7.3	0

Some of the simulation results are summarized in Table II and in Figures 1 through 5. We note that the plots of state estimates for cases 2 and 4 (refer to Table I)

TABLE II. Summary of Monte Carlo simulation results.

Case	L-N Simulation		B-N-L Simulation				Kalman Filter			
			M = 10		M = 20		Simulation		Computed	
	Mix.	Norm.	Mix.	Norm.	Mix.	Norm.	Mix.	Norm.	Mix.	Norm.
1&3	0.0291	0.0271	0.0237	0.0222	0.0227	0.0216	0.1148	0.0941	0.0956	0.0864
2&4	0.2387	0.2241	0.1277	0.1199	0.1100	0.1086	1.0372	0.8633	0.5311	0.8826

are not shown since the high variance plant and measurement noise created rather noisy looking estimates. As a matter of fact, cases 2 and 4 (with $r \doteq 1$) represent the "high Q" KF where we tend to ignore the dynamic state model and place emphasis on the measurements. For the "low Q" cases ($r = 0.01$), Fig. 3 is a composite plot of the state estimates of the L-N, B-N-L (with M = 20) and the KF in mixture noise, illustrating the

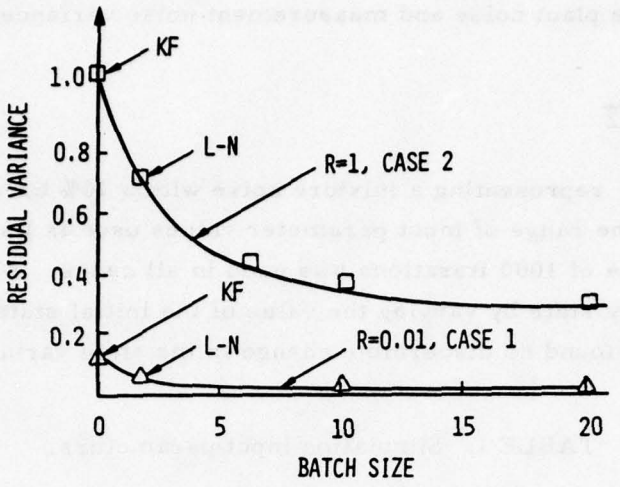


Fig. 1. Residual variance comparisons for the L-N and B-N-L robustized and standard Kalman filters.

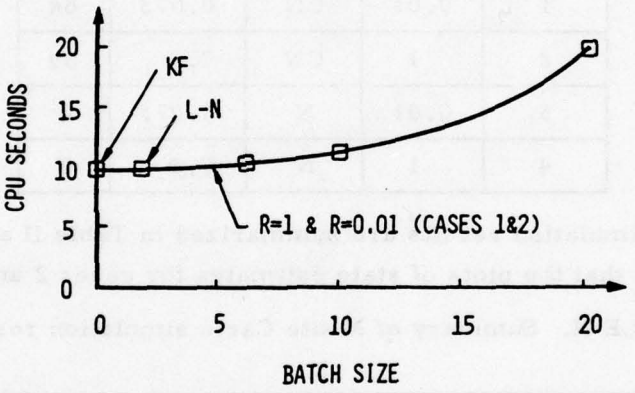


Fig. 2. Computation time comparison for the L-N and B-N-L robustized and standard Kalman filters.

robustness of the B-N-L approach via the MWWNS. Figures 4 and 5 are the individual plots of the state estimates in mixture noise for the KF, L-N, B-N-L with M = 10, B-N-L with M = 20 filters, respectively.

It follows by inspection of Table II that the B-N-L filter outperforms the L-N filter in all cases which is illustrated in Fig. 1 by comparing the Monte Carlo state variances. We also see from Fig. 1 that the B-N-L filter state variance reaches almost asymptotic value with a batch size between 6 and 10 and the per cent difference (for case 1, "low Q" in mixture noise) between the state variances for M = 10 and M = 20 is 4%. For case 3 ("low Q" in Gaussian noise) the comparable difference is 2.8%.

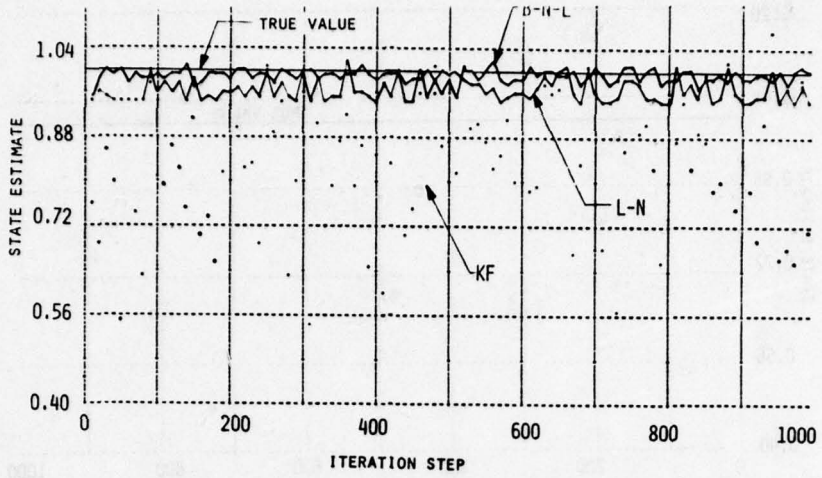


Fig. 3. A composite comparison of state estimates for the L-N and B-N-L robustized and standard Kalman filters in mixture noise.

These results compare within 10 to 1 to the theoretical value of 0.37%. However, the above results are in good agreement with theory when we compare them with the variation between the computed and Monte Carlo state variances for the KF of 2.2% for case 3 and 95% for case 2, respectively. The robustness property is also apparent from Table II for the B-N-L filter. The difference between cases 1 and 3 for the L-N filter is 9% while the corresponding difference for the B-N-L filter is only 3% or the latter approach is more robust.

A comparison of the relative computation time for the robust and standard Kalman filters on the IBM 360/67 time-sharing computer is shown in Figure 2. We see that while batch processing requires approximately 1.2 to 2 longer time than KF or the L-N filter, its robustness properties and ease of implementation outweigh its added cost in running time (compare Figures 1 and 2).

In summary, the Monte Carlo simulation of the robust L-N, B-N-L and Kalman filters substantiated the theoretical results and demonstrated the advantages of the batch-nonlinear-linear filter structure using the MWWNS.

Joint Services Technical Advisory Committee
F44620-74-C-0056

I. Kadar and L. Kurz

Grumman Aerospace Corporation

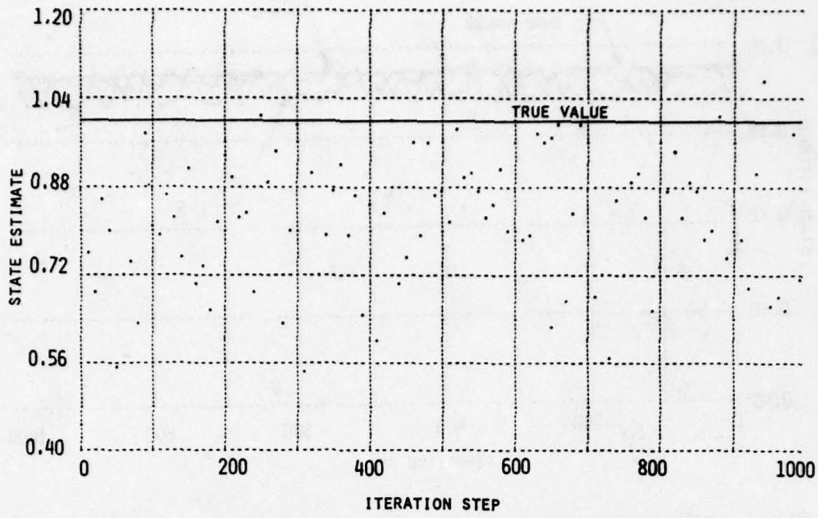


Fig. 4. State estimate for the standard Kalman filter in mixture noise.

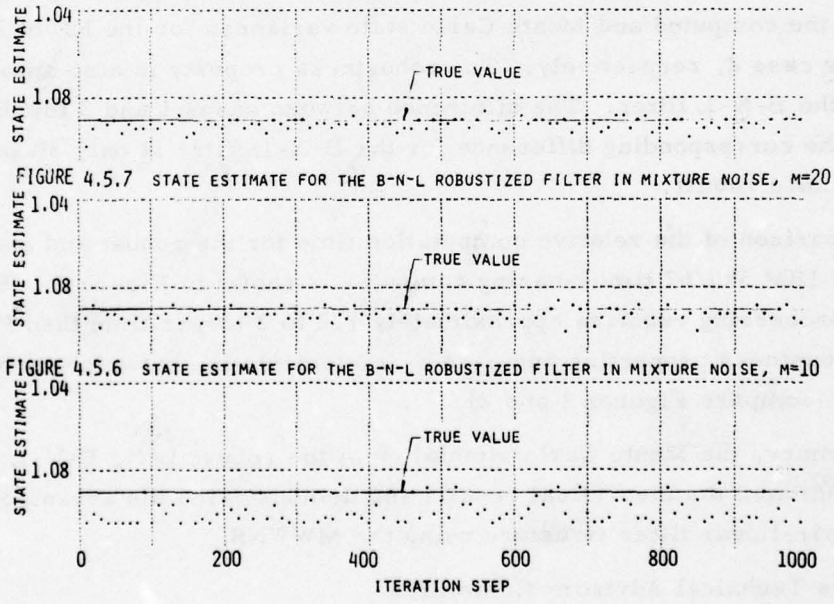


Fig. 5. State estimate for the L-N robustized filter in mixture noise.

REFERENCES

1. P.J. Huber, "Robust Estimation of a Location Parameter," Ann. Math. Stat., Vol. 35 (1964).
2. R.D. Martin, "Robust Estimation of Signal Amplitude," IEEE Trans. on Information Theory, Vol. IT-18, No. 5 (September 1972).
3. J. Evans, P. Kersten and L. Kurz, "Robustized Recursive Estimation with Applications," Information Sciences, Vol. 11, No. 1, pp. 69-92 (1976).
4. P. Kersten and L. Kurz, "Robustized Vector Robbins-Monro Algorithm with Applications to M-Interval Detection," Information Sciences, Vol. 11, No. 2, pp. 121-140 (1976).
5. J. Evans, "Robust Stochastic Approximation Procedures Using Nonparametric Statistics in Unspecified Noise Environments," Ph.D. Thesis, New York University, (October 1973).
6. A. Segall, "Recursive Estimation from Discrete-Time Point Processes," IEEE Trans. on Information Theory, Vol. IT-22, No. 4 (July 1976).
7. L.A. Jaeckel, "Robust Estimates of Location: Symmetry and Asymmetric Contamination," Ann. Math. Stat., Vol. 42, No. 3 (1971).
8. C.J. Masreliez and R.D. Martin, "Robust Bayesian Estimation for the Linear Model and Robustizing the Kalman Filter," Eighth Annual Princeton Conf. on Information Sciences and Systems, pp. 488-492 (1973).
9. M.W. Woinsky and L. Kurz, "A Noncoherent Detector with Improved Immunity to Fading," Eascon Conf. Record (1968).
10. R.D. Martin and C.J. Masreliez, "Robust Estimation via Stochastic Approximation," IEEE Trans. on Information Theory, Vol. IT-21, No. 3 (May 1975).
11. M. Chernoff and I.R. Savage, "Asymptotic Normality and Efficiency of Certain Nonparametric Test Statistics," Ann. Math. Stat., Vol. 29 (1958).
12. J. Hajek, "Asymptotic Normality of Simple Linear Rank Statistics under Alternatives," Ann. Math. Stat., Vol. 39, No. 2 (1968).
13. L. Kurz, "Nonparametric Detectors Based on Partition Tests," Polytech. Inst. of New York, EE 704 Supplementary Notes (September 1975).
14. M.B. Mann and D.R. Whitney, "On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other," Ann. Math. Stat., Vol. 18 (1946).
15. E. Fix and L.J. Hodges, Jr., "Significance Probabilities of the Wilcoxon Test," Ann. Math. Stat., Vol. 26 (1955).
16. A.P. Sage and J.L. Melsa, "Estimation Theory with Applications to Communications and Control," (McGraw-Hill, 1971).
17. Y.C. Ho, "On the Stochastic Approximation Method and Optimal Filtering Theory," J. Math. Analysis and Applications, Vol. 6 (1962).
18. I. Kadar, "Robustized Estimation and Image Processing," Ph.D. Dissertation, Polytech. Inst. of New York (June 1977).
19. J.L. Hodges and E.L. Lehmann, "Estimates of Location Based on Rank Tests," Ann. Math. Stat., Vol. 34 (1963).

ROBUSTIZED SCALAR FORM OF GLADYSHEV'S THEOREM WITH APPLICATIONS TO NONLINEAR SYSTEMS

I. Kadar and L. Kurz

Gladyshev[1] in 1965 gave a simplified proof of the scalar Robbins-Monro Stochastic Approximation[2] (RMSA) procedure permitting the restrictions imposed on the unknown regression function to be weakened somewhat, extended convergence results in the spirit of Sacks[3] and proposed a theorem (without proof) which is an extension of RMSA to least squares.

The emphasis in this report is on the unproven portion of Gladyshev's theorem, robustizing it along the lines suggested by Evans, Kersten and Kurz[4] and application of the results to problems of current interest, i.e., an interferometer used for spacecraft altitude control which measures angle of arrival (AOA), a correlator receiver, a tone-ranging system which measures time of arrival (TDA), etc.

Essentially, the theorem is a statement of the orthogonality principle in the RMSA framework which is termed Stochastic Approximation Minimum Variance Least Squares (SAMVLS). The theorem is robustized via the batch-nonlinear-linear (B-N-L) and linear-nonlinear (L-N) approaches.

A. RMSA Extension to Least Squares (SAMVLS) and Small Sample Theory

The following theorem establishes the asymptotic convergence of RMSA to minimum variance least squares:

Theorem: Let (ξ, η), (ξ_1, η_1), $\cdots$, be a sequence of two-dimensional real random vectors, $E[\xi^4] < \infty$, $E[\eta^4] < \infty$. Further, let $x_1, x_2, \cdots$, be a random sequence in which x_1 is an arbitrary real random variable and $x_2, x_3, \cdots$, are determined from

$$x_{n+1} = x_n - (A/n)(\xi_n x_n - \eta_n)\xi_n$$

where $A(\cdot) > 0$, then the sequence $\{x_n\}$ converges w.p.1 to a value θ which turns the expression $E[(\xi\theta - \eta)^2]$ into a minimum. If, in addition $AE[\xi^2] > \frac{1}{2}$, the sequence $\sqrt{n}(x_n - \theta)$ is asymptotically normal with mean zero and variance

$$\sigma^2 = A^2(2AE[\xi^2] - 1)^{-1} E\left[\theta\xi^2 - \theta E[\xi^2] - \xi\eta + E[\xi\eta]\right]$$

Proof:

If we recognize that

$$M(x_n; \theta) = E[Y(x_n; \theta)] = E[(\xi_n x_n - \eta_n)\xi_n]$$

then the proof follows a similar procedure as in Theorem B1 of Reference 12. Because of the length and complexity, the details of the proof are omitted.

Since in practice one always deals with a finite sample size, the question naturally arises as to how well asymptotic theory approximates the true situation. We examine Gladyshev's extension of RMSA to least squares in the small sample case (with constant gain coefficients). This illustrates the significance and the need for adaptive gain coefficients in terms of efficiency of the procedure.

With the linear regression function of the form $M(x_n; \theta) = \alpha_{in}(x_n - \theta) + b_n$ where $\alpha_{in} = E[\xi^2]$, $b_n = -E[\xi_n \eta_n]$, $Y(x_n; \theta) = M(x_n; \theta) + Z(x_n)$ and $Z(x_n) = \xi_n^2 x_n - \xi_n \eta_n - E[\xi^2] x_n + b_n$, we let $b_n = 0$ which merely shifts the intercept of the regression function. Furthermore, we introduce a sequence of i.i.d.r.v. ε_n satisfying

$$E[\varepsilon_n] = 0, \quad E[\varepsilon_n^2] = 1 = \operatorname{var} \varepsilon_n$$

$$Z(x_n) = \varepsilon_n \rho$$

such that

$$E[Z(x_n)] = 0 \quad \text{and} \quad \lim_{x_n \to \theta} E[Z^2(x_n)] = \rho^2$$

as required in the Theorem.

We estimate the root θ of $M(x)$ by the RMSA method, i.e., choose x_1 arbitrarily and $x_2, x_3, \cdots$, are calculated from

$$x_{n+1} = x_n - a_n[\alpha_1(x_n - \theta) + \varepsilon_n \rho]$$

Next, the residuals

$$(x_n - \theta) - a_n \alpha_{in}(x_n - \theta) - a_n \varepsilon_n \rho$$

are formed. Let $V_n = x_n - \theta$ and $a_n = a/n\alpha_{in}$. It follows that

$$V_{n+1} = V_n(1 - a/n) - (a/n)(\rho/\alpha_{in})\varepsilon_n$$

Let $V_1 = \Delta$, and iterating back, it can be easily shown that[5]

$$V_n = \Delta \prod_{k=1}^{n-1} (1 - \frac{a}{n}) - \rho \sum_{i=1}^{n-1} \left[\frac{a}{i} \prod_{k=i+1}^{n-1} (1 - \frac{a}{k}) \frac{\varepsilon_i}{\alpha_{1i}} \right] \tag{1}$$

By taking expectations, one obtains

$$E[V_n] = E[(x_n - \theta)] = \Delta \prod_{k=1}^{n-1} (1 - \frac{a}{k}) \tag{2}$$

which is the expected bias at step n. It gives a measure how far down the sequence an inappropriate choice of V_1 biases the results.

The variance of V_n is given by

$$\mathrm{Var}(V_n) = \mathrm{Var}(x_n) = \rho^2 \sum_{i=1}^{n-1} \left[\frac{a}{i} \prod_{k=i+1}^{n-1} \left(1 - \frac{a}{k}\right) \right] \frac{1}{\alpha_{1i}^2} \tag{3}$$

Thus, Eqs. (1)-(3) give the exact bias and variance of the estimate of x_n at any step n of the estimation process. One can easily show, by letting $(1 - a/k) \approx e^{-a/k}$, that

$$\lim_{n \to \infty} E[V_n] = 0 \tag{4}$$

and

$$\lim_{n \to \infty} \mathrm{Var}(V_n) = \rho^2 \alpha^2 / [n \alpha_1^2 (2a - 1)] \tag{5}$$

Now if we redefine $a_n = A/n$, where $A = [a/E(\xi^2)]$, we obtain the desired form of the result

$$\lim_{n \to \infty} \mathrm{Var}(V_n) = \frac{\rho^2 A^2}{n[2AE(\xi^2) - 1]}, \quad AE[\xi^2] > \frac{1}{2} \tag{6}$$

The results of Eqs. (4), (5) and (6) show that under certain conditions on $AE[\xi^2]$, the biases arising from bad initial guesses rapidly tend to zero with increasing sample size, and the expected squared error of x_n is of order $1/n$.

Sacks[3] showed under general conditions, with $a_n = (A/n)$, x_n is asymptotically normal with variance given by Equation (6). A suitable choice for A is to minimize Equation (6). This leads to choosing $A = E^{-1}(\xi^2)$ where

$$\lim_{n \to \infty} \mathrm{Var}(V_n) = \rho^2 / n E(\xi^2) \tag{7}$$

asymptotically. If we assume that $Y(x; \theta)$ is normally distributed with mean $E[\xi^2](x - \theta)$ and variance ρ^2, we have

$$1/E[\partial \log f(x; \theta)/\partial \theta] = \rho^2 / E[\xi^2] \tag{8}$$

and the Cramer-Rao lower bound on the variance of estimates of θ is Equation (8). The amount of Fisher information for the above case in a sample size n is given by

$$n I[f; \theta=0] \equiv n \int_{-\infty}^{\infty} \left(\frac{\partial \log f}{\partial \theta}\right)^2 f(x)\, dx \Bigg|_{\theta=0}$$

$$= n \int_{-\infty}^{\infty} \left(\frac{f'(x)}{f(x)}\right)^2 f(x)\, dx = n \frac{E(\xi^2)}{\rho^2}$$

or the Fisher information per observation is $E(\xi^2)/\rho^2$.

A more general result holds from the following considerations. The Gladyshev form of the estimator can be fully efficient (asymptotically) if $A = E^{-1}(\xi^2)$. The practical difficulty arises that $E^{-1}(\xi^2)$ is unknown and must be guessed or estimated. This is precisely why adaptive gain coefficients are introduced to estimate the slope (near the root) of the regression function at every step. However, the efficiency of this process is weakly dependent on the correct choice of $E^{-1}(\xi^2)$. The asymptotic efficiency is Eq. (7) divided by Eq. (6) which yields

$$\text{Eff} \equiv (2A E(\xi^2) - 1)/A^2 E^2(\xi), \quad A E(\xi^2) > \frac{1}{2} \tag{9}$$

A plot of Eq. (9) shows a broad maximum and a tendency toward zero efficiency at $A E(\xi^2) = 1/2$ in a very gradual manner. This has been substantiated by Hodges and Lehmann[6] who have shown that for $A E(\xi^2) = 1/2$ the asymptotic variance is given by

$$\text{Var}(V_n) = \frac{\rho^2 \log n}{4n E^2(\xi^2)}$$

and the efficiency becomes $4/(\log n + \gamma)$ where γ is the Euler's constant.

In terms of the requirements for the estimator of the adaptive gain coefficient[4] $A_n(x_1, \cdots, x_n)$, the above result points out that an exact estimate of $E^{-1}(\xi^2)$ (in general, the slope of the regression function) is not needed at every step because of the broadness of the efficiency curve. The conditions required for the estimator of the adaptive gain coefficient, $A_n(\cdot)$ are: $E[A(\cdot)] \to A$, $E|A^2 - E(A_n^2(\cdot))| \to 0$ and $E[A_n^2(\cdot)] \to A$ as $n \to \infty$; in addition, A in the neighborhood of $E^{-1}(\xi^2)$ will give high efficiency at each step.

B. Minimax Theory Applied to Rank Tests and SAMVLS

In robustizing SAMVLS two approaches are possible: Huber's[7] light limiter influence function (LLIF) or M-estimators and rank statistic preprocessor,[4] or R-estimators. In this section, the mathematical background which permits meaningful comparison of these two approaches is presented.

Let $x_1, x_2, \cdots, x_n$ be i.i.d.r.v. with c.d.f. $F(x - \theta)$ and p.d.f. $f(x - \theta)$. Huber[7] proposed a maximum likelihood estimator of θ called an M-estimate, which satisfies

$$\sum_{i=1}^{w} \psi(x_i - M) = 0 \quad ,$$

where $\psi(x)$ is such that $\psi(-x) = -\psi(x)$. If $\psi(x)$ is monotonic, M is essentially uniquely determined. Huber[7] has shown that the M-estimator is asymptotically normal with mean θ and asymptotic variance of $n^{1/2}M$.

$$V_M = \frac{\int \psi^2(x) f(x)\,dx}{[\int \psi'(x) f(x)\,dx]^2} \geq \frac{I}{\int (f'/f)^2 f\,dx}$$

This theory is useful in the study of mixture distribution noise of the form $F = (1-\epsilon)G + \epsilon H$, where $0 \leq \epsilon \leq 1$ is a fixed number, G is a fixed and symmetric distribution and H is a variable c.d.f. Huber demonstrated that there exists a family of distributions which includes the mixture distribution for which there exists a saddle point $V_M(\psi_o, F_o)$ satisfying

$$V_M(\psi, F_o) \leq V_M(\psi_o, F_o) = I(F_o) \leq V_M(\psi_o, F)$$

The least favorable density, f_o, is Gaussian in the middle and double-exponential on the tails. The M-estimator is the corresponding maximum likelihood estimator with $\psi_o = -f_o'/f_o$. The associated ψ_o is given by the LLIF $\psi_o(x) = \ell_k(x)$ which satisfies the conditions of the minimax result.

Jackel[8] has shown that the minimax theory established by Huber also applies to the estimators derived from rank tests, as well as to estimators based upon linear combination of order statistics with the asymptotic variances of the three classes of estimators being identical. With rank tests of the Chernoff-Savage form

$$T_N = \sum_{i=1}^{N} a_{N_i} Z_{N_i}$$

with scores $a_{N_i} = J_N(i/N)$, we define the R-estimator as the solution of $J_N(R) = 0$ under the hypothesis with $m = n$, $N = 2w$. The asymptotic variance of $n^{1/2}R$ is then

$$V_R(\psi, F) = \frac{\int J^2(t)\,dt}{[\int \frac{d}{dx}\{J[F(x)]\}\, f(x)\,dx]^2} \tag{10}$$

and with $t = F(x)$, $J(t) = \psi(x)$, $V_M(F) = V_R(F)$ asymptotically. We note that the numerator of V_R does not depend on F unlike the numerator of V_M. This indicates that the R-estimator will more likely be more robust than the M-estimator. This observation will be further supported in subsequent discussion and simulations.

Applying the minimax theory for the M-estimator, the R-estimator corresponding to the M-estimator is defined by

$$J_o(t) = \psi_o(x)$$

with F_o defined by

$$\begin{aligned} f_o(x) &= (1-\epsilon)\, g(-x_o) \exp[k(x + x_o)] && \text{for} \quad x \leq -x_o \\ &= (1-\epsilon)\, g(x) && \text{for} \quad -x_o < x < x_o \\ &= (1-\epsilon)\, g(x_o) \exp[-k(x - x_o)] && \text{for} \quad x \geq -x_o \end{aligned}$$

for some X_o and k depending on G and ϵ. $\psi_o(X) = -f_o'(X)/f_o(X)$, a generalized LLIF, is monotonic, and for $X < -X_o$ and $X > X_o$ is a constant, and

$$V_R(\psi, F) \leq V_R(\psi_o, F_o)$$

This result, along with the minimax bound for the M-estimator, allows us to obtain lower bounds for the asymptotic variance of the robust SAMVLS. If we define a general nonlinear odd-symmetric robustizing transformation, T satisfying the requirements of the theorem in Section A, which may represent either a nonparametric rank statistic or the LLIF, then the asymptotic variance of the robust SAMVLS is

$$V_{SA} = \frac{\text{Var } T(\cdot)}{n\alpha_1^2}\left[\frac{A^2\alpha_1^2}{2A\alpha_1 - 1}\right] \geq \frac{\text{Var } T(\cdot)}{n\alpha_1^2}$$

where

$$V_T(\ ) = \frac{\text{Var } T(\cdot)}{n\alpha_1^2} \leq V_T(\psi_o, F_o)$$

and

$$\alpha_1^2 = \{E_F[T'(\cdot)]\}^2$$

Two types of nonparametric statistics: a two-sample Mann-Whitney-Wilcoxon (MWWNS) of the form

$$W = \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}\text{sgn}(x_i - y_j)$$

and Wilcoxon one-sample symmetrical (WSRNS) of the form

$$S = \sum_{i>j}^{m}\text{sgn}(x_i - x_j) + \sum_{j=1}^{m}\text{sgn } x_j$$

are particularly useful as B-N-L robustizers of SAMVLS. While requiring storage, the batch nonparametric statistic preprocessed transformation achieves, with a moderate sample size of m, approximate asymptotic normality, with a constant variance, independent on the underlying CDF.

Unfortunately, this is not the case for the light limiter influence function robustizing transformation, where the limiter merely removes outliers. Furthermore, the asymptotic normality and robustness properties of the corresponding M-estimator are only applicable asymptotically since in the case of the LLIF transformation there is no batch preprocessing corresponding to the MWWNS or the WSRNS.

C. A Comparison of Robustizing Methods with Applications to Nonlinear Systems

In the previous sections we have shown that the nonparametric rank statistics, (i.e., MWWNS and WSRNS) and the LLIF robustizing transformations applied to RMSA are asymptotically equivalent. The main differences are in the moderate sample asymptotic behavior and in the rate of convergence.

In order to compare the performance of the robust SAMVLS under different robustizing transformations, we assume that the noise is described by the mixture distribution model. This allows us to select the parameters of the robustized recursion under the nominal CDF's and determine the degree of robustness in the presence of contamination. One-dimensional applications oriented examples drawn from practical systems illustrate the above approach.

To determine time-of-arrival (TOA) or to measure the angle-of-arrival (AOA) and in other application areas,[4] many sensors, navigation and communication systems utilize correlator-discriminators and phase detectors. These devices compare the difference between possibly nonlinear functions of two signals in additive noise to measure the unknown signal parameters. One signal is usually taken as the reference (i.e., known) and the parameter measured (i.e., unknown) is derived with respect to that reference.

The generic characteristics of these devices is usually nonlinear, their region of operation is usually restricted to a linear portion of the characteristics. The output of these devices is a noisy error signal which is linearly related to the difference between the two signal parameters which drives a control loop to adjust the reference signal with respect to the incoming signal in a direction to drive the error (i.e., the difference) to zero.

Having sampled values available of such systems in a digital implementation, we have the exact framework, with the appropriate identification of terms, for RMSA applied to least squares (SAMVLS). Three examples for such systems are given; an interferometer used in spacecraft attitude control to measure AOA (see Ref. 9), the discriminator portion of a correlation receiver used to measure relative TOA (see Ref. 10), and a CW phase-comparison ranging system (see Reference 11). We assume that the systems under consideration are bandlimited and are narrow band with respect to some center frequency. The output of these systems is sampled for digital processing at a rate slow enough to have negligible cross-correlation between the signal and noise samples, i.e., the samples are i.i.d.

In the systems under consideration, the noise is modelled as a stationary narrow band process. The probability density function of the phase of the sinusoidal signal plus

narrow band noise is given by[12]

$$f(\gamma) = \frac{\exp(-A^2\ 2R(o))}{2\pi} + \frac{A\cos(\gamma)\exp[-A^2\sin^2(\gamma)\ 2R(o)]}{2\sqrt{2\pi R(o)}}\left[1 + \text{erf}\frac{A\cos(\gamma)}{R(o)}\right]$$

for $|\gamma| < \pi$ and zero elsewhere, where A is the signal amplitude, $R(o) = \sigma^2$ noise. If the input signal-to-noise ratio (SNR) is high in both channels (e.g., interferometer), SNR = $A^2/2R(o)$ and we limit the range of angle measurement to $|\gamma| < 5^o$ (i.e., $\cos\gamma \simeq 1$, $\sin\gamma \simeq \gamma$), then the expression for $f(\gamma)$ reduces to

$$f(\gamma) = \frac{\exp(-s^2\gamma^2)}{\sqrt{\pi/s^2}},\ s^2 \gg 1,\ |\gamma| < 5^o$$

where, for notational convenience, we let S^2 = SNR, and $f(\gamma)$ is Gaussian with zero mean and variance $1/2S^2$. The sinusoidal signal plus narrow band noise appearing at the inputs to the phase detector are multiplied and the difference between their phases (having removed the sum frequency terms) is given by $\phi = \alpha - \beta + v$ where, v is Gaussian with mean zero and variance σ^2.

Consider the SAMVLS to adjust $\alpha = \beta$ to reorient a spacecraft or to measure relative range, with sequential measurements available. In the case of the correlator-discriminator this is exactly equivalent to aligning the epoch of the replica code with the incoming code to determine relative TOA. For the phase comparison ranging system, this is equivalent to fine alignment of the reference tone phase to match the incoming tone phase. In terms of the SAMVLS, these operations can be defined with $\xi = 1$ and $\gamma(\alpha_n; \theta) = \alpha_n - \theta + v_n$. The SAMVL recursion is then of the form

$$\alpha_{n+1} = \alpha_n - (A/n)[\gamma(\alpha_n; \theta)] \tag{11}$$

To robustize Eq. (11) one introduces either batch preprocessing and a nonparametric rank statistic or the LLIF. The output of either combination is then applied to $\gamma(\alpha_n; \theta)$. The MWWNS robustized SAMVL recursion is of the form

$$\alpha_{n+1} = \alpha_n - \frac{A_w}{n}\left[\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\text{sgn}(\alpha_{[m(n-1)+i]} - \theta_{[m(n-1)+j]} + V_{[m(n-1)+i]})\right]$$

The WSRNS robustized SAMVL recursion is of the form with

$$\phi_i = \alpha_{n+i} - \theta_{n+i} + V_{n+i}$$

$$\alpha_{n+1} = \alpha_n - (A_s/n)S^m(\phi)$$

where

$$S^m(\ ) = \sum_{i>j}^{m} \operatorname{sgn}(X_i + X_j) + \sum_{i=1}^{m} \operatorname{sgn}(X_i)$$

The LLIF robustized SAMVLS recursion is of the form

$$\alpha_{n+1} = \alpha_n - (A_L/n)\, \ell_k \left[\alpha_n - \theta_n + V_n\right]$$

where

$$\ell_k(\ ) = \begin{cases} x\ , & |x| < k \\ \operatorname{sgn}(x), & |x| \geq k \end{cases}$$

To evaluate the optimum gain coefficients for the above recursions, we have to find the respective regression functions in a form which satisfies theorem 1 of Ref. 4 and evaluate the derivative of the regression function at the root.

In reference to the theorem of Section A, the parameters of the MWWNS robustized recursion are with

$$W^m(\ ) = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \operatorname{sgn}(X_i - Y_j)$$

and define

$$U^m(\ ) = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} U(X_i - Y_j)$$

such that $W = 2U - 1$; then it can be shown that

$$E\,W = \left[2 \int_o^{\infty} f_{x-y}(z/\alpha\ , \theta)\, dz - 1\right]$$

where $f_{x-y}(\ )$ is the pdf of difference between independent variables $(X_i - Y_j)$.

The reciprocal of the optimum gain coefficient $\alpha_1 = 1/A$ is given by

$$E\,W = \alpha_1^W(\alpha - \theta) + O(|\alpha - \theta|^2)$$

where

$$\alpha_1^W = 2 \frac{\partial}{\partial \alpha} \int_o^{\infty} f_{x-y}(z/\alpha, \theta)\, dz \Bigg|_{\alpha=\theta} = 2 f_{v_1 - v_2}(o)$$

and

$$f_{v_1 - v_2}(o) = f_{v_1 + v_2}(o) = 1/\sqrt{4\pi}\sigma_v$$

since V_n is an i.i.d. sequence of symmetric Gaussian r.v.'s. The asymptotic variance of the WMWNS robustized recursion is given by

$$V_{SA}^{W} = V_R^W \left[\frac{A^2/4\pi\sigma_v^2}{4A/(\sqrt{4\pi}\sigma_v) - 1} \right] \geq V_R^W$$

where

$$V_R^W = \frac{(2m+1)(1/3m^2)}{n(1/4\pi\sigma_v^2)} \leq V_R^W(\psi_o, F_o)$$

For a WSRNS robustized SAMVLS proceeding in a manner analogous to the MWWNS, details of which are covered in Ref. 4

$$\alpha_1^s = 2\left[m(m-1) f_{v_1+v_2}(o) + m f_v(o)\right]$$

where under the nominal Gaussian CDF for v,

$$f_{v_1+v_2}(o) = \frac{1}{\sqrt{4\pi}\sigma_v} \quad \text{and} \quad f_v(o) = \frac{1}{\sqrt{2\pi}\sigma_v} .$$

The asymptotic variance of the WSRNS robustized SAMVLS is given by

$$V_{SA}^S = V_R^S[A^2(\alpha_1^S)^2 / (2A\alpha_1^S - 1)] \geq V_R^S$$

where

$$V_R^S = [(1/24)(m+1)(2m+1)/n(\alpha_1^S)^2] \leq V_R^S(\psi_o, F_o)$$

In the LLIF robustized SAMVLS, in reference to the theorem in Section A, the salient parameters of the robust recursion are with $\eta = \theta + v$

$$E\,\ell_k\, Y(\alpha, \theta') = \int_{-\infty}^{\infty} \ell_k(\alpha - \theta') f_v(\alpha - \theta)\, d\alpha$$

and

$$\alpha_1^L = \int_{-k}^{k} f_v(x)\, dx = [1 - 2\,\mathrm{erf}(k)]$$

if $f_v(x)$ is Gaussian. The asymptotic variance of the LLIF robustized SAMVLS is given by

$$V_{SA} = V_L[A^2(\alpha_1^L)^2/(2A\alpha_1^L - 1)] \geq V_L$$

where

$$V_L = \frac{\int_{-\infty}^{\infty} \ell_k^2(x) f_v(x)\, dx}{n(\alpha_1^L)^2} \leq V_L(\psi_o, F_o)$$

A series of Monte-Carlo simulations were implemented to evaluate the performance of each robustized SAMVLS recursion, as well as the performance of unrobustized SAMVLS in Gaussian, N(0, 1) and in mixture noise. Two models were used for mixture noise. In the first model the nominal Gaussian noise N(0, 1) was contaminated with a 10% high-variance Gaussian burst-noise component, i.e., 0.9 N(0, 1) + 0.1 N(0, 8). In the second case, the Gaussian burst-noise component was replaced with 2% Cauchy, CN(0, 64), i.e., 0.98 N(0, 1) + 0.02 CN(0, 64). The Monte Carlo noise generator characteristics are summarized in Table I for a step size of 1000 and 10,000 iterations.

TABLE I. Monte Carlo noise generator characteristics.

Measurement Noise Simulated	Mean	Standard Deviation	Minimum	Maximum	Number of Samples
1 +N(0, 1)	1.01319	1.00455	-2.89972	4.8185	10,000
1 +0.9N(0, 1) + 0.1N(0, 8)	-3.78762	0.90516	-7.30975	0.6424	10,000
1 +0.98N(0, 1) + 0.02CN(0, 64)	2.03288	79.29695	-1251.17603	5079.6875	10,000
1 +N(0, 1)	1.06148	0.97842	-2.89972	4.5472	1,000
1 +0.9N(0, 1) + 0.1N(0, 8)	-3.73960	0.89117	-7.30975	0.6424	1,000
1 +0.98N(90, 1) + 0.02CN(0, 64)	1.04001	23.10217	-388.38037	247.1582	1,000

Note that even in the nominal Gaussian case, there is a slight positive bias in the noise generator. For the MWWNS and the WSRNS, a batch size of m = 10 was selected, which allowed comparison of the convergence rates between the two batch-integer-rank preprocessed SAMVLS.

First, the simulations were computed for the Gaussian mixture case with a 10% burst-noise component. The simulation results for a step-size of 1000 iterations are summarized in Table II and in Figures 1 and 2. Note that the ratios of the sample variances in Table I between the MWWNS robustized and WSRNS robustized SAMVLS is 1.898, which is very close to the theoretically predicted value of ~2. Figure 1 shows

TABLE II. Monte Carlo simulation comparison summary. (1,000 samples), scalar SAMVLS in mixture noise.

Robustizing Method	Mean	Variance	Minimum	Maximum	Batch Size
MWWNS (Two Sample)	1.01438	0.07239	-1.32851	1.15423	10
WSRNS (One Sample)	0.96122	0.03815	-0.55250	1.03668	10
LLIF (k = 0.317)	0.90488	0.00344	0.31700	0.94291	---
Unrobustized	1.12588	0.00404	0.64237	1.48352	---

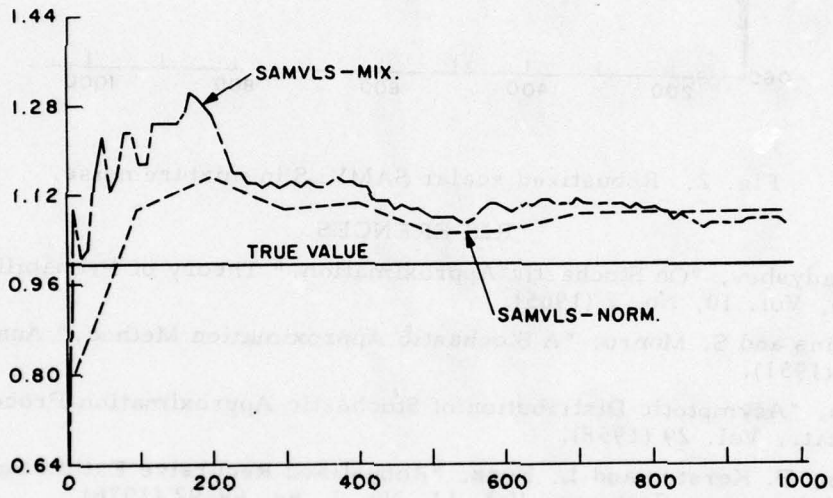


Fig. 1. Unrobustized scalar SAMVLS in Gaussian and mixture noise.

the performance of SAMVLS without any robustizing transformation in nominal Gaussian and mixture noise. Without robustizing, the convergence of the recursions is slow.

The performance of the robustized scalar SAMVLS recursion is shown in Figure 2. Note that both the WSRNS and MWWNS reach true values after 200-300 iterations. However, the LLIF is biased, as shown in Figure 2. While converging fast with a small sample variance, it converges to a biased value, which depends on the nonlinearity and the bias cannot be predicted. This demonstrates the drawback of the L-N approach via the LLIF. At the same time, the WSRNS and the MWWNS converge fast to the true parameter values.

Joint Services Technical Advisory Committee
F44620-74-C-0056

I. Kadar and L. Kurz
Grumman Aerospace Corporation

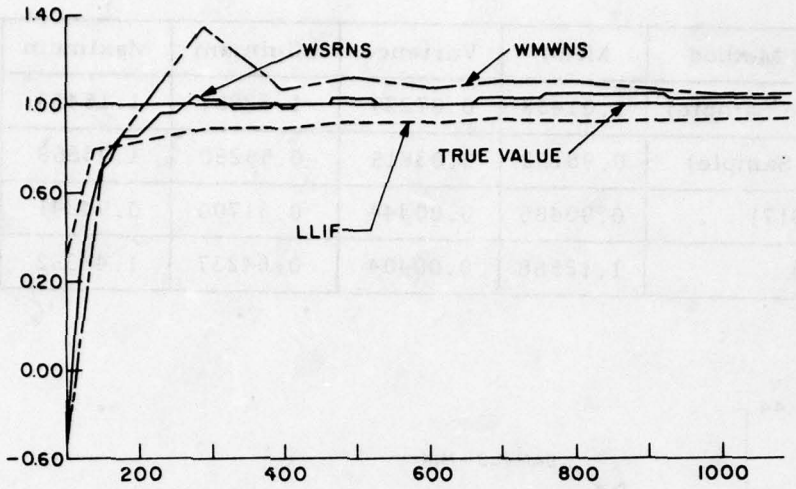


Fig. 2. Robustized scalar SAMVLS in mixture noise.

REFERENCES

1. E.G. Gladyshev, "On Stochastic Approximation," Theory of Probability and Its Application, Vol. 10, No. 2 (1965).
2. H. Robbins and S. Monro, "A Stochastic Approximation Method," Ann. Math. Stat., Vol. 22 (1951).
3. J. Sacks, "Asymptotic Distribution of Stochastic Approximation Procedures," Ann. Math. Stat., Vol. 29 (1958).
4. J. Evans, P. Kersten and L. Kurz, "Robustized Recursive Estimation with Applications," Information Sciences, Vol. 11, No. 1, pp. 69-92 (1976).
5. M.T. Wasan, "Stochastic Approximation," (Cambridge U. Press, 1969).
6. J.L. Hodges and E.L. Lehmann, "Two Approximations to the Robbins-Monro Process," Proc. 3rd Berkeley Symp. Math. Stat. and Prob. (1959).
7. P.J. Huber, "Robust Estimation of a Location Parameter," Ann. Math. Stat., Vol. 35 (1964).
8. L.A. Jaeckel, "Robust Estimates of Location: Symmetry and Asymmetric Contamination," Ann. Math. Stat., Vol. 42, No. 3 (1971).
9. A. Segall, "Recursive Estimation from Discrete-Time Point Processes," IEEE Trans. on Information Theory, Vol. IT-22, No. 4 (July 1976).
10. I. Kadar, "Performance of Correlation Receivers in an Interference Environment," Proc. IEEE 1975 National Aerospace and Electronics Conference, pp. 355-362 (1975).
11. I. Kadar, "An Analysis of Helicopter Rotor Modulation Interference," IEEE Trans. on Aerospace and Electronics Systems, Vol. AES-9, No. 3, pp. 434-444 (May 1973).
12. J.J. Courtney, Jr. and I. Kadar, "A Ground Simulation of a Satellite Navigation System," Progress in Astronautics and Aeronautics-Communication Satellite Developments: Systems, Vol. 41, G.E. LaVean and W.G. Schmidt, Eds. (MIT Press, 1976), pp. 199-243.

SIGNAL PROCESSING

A. Papoulis

Current work in Signal Processing involves one and two-dimensional deterministic and random signals. The following are brief descriptions of recent results on various aspects of the program. Details can be found in the stated references.

A. The Two-to-One Rule in Data Smoothing[1]

If a signal is estimated by a weighted average of the data in the interval (t-c, t+c), then the variance σ^2 of the estimate decreases, but its bias b increases, with increasing c. It is shown that in high accuracy estimates, the mean-square error e is minimum if c is such that $\sigma = 2b$, regardless of the form h(t) of the smoothing weight. Furthermore, the resulting e_m is minimum if h(t) is a truncated parabola.

B. Generalized Sampling Expansion[2]

It is shown that a bandlimited function f(t) is uniquely determined in terms of the samples $g_k(nT)$ of the responses $g_k(t)$ of m linear systems with input f(t), sampled at 1/m the Nyquist rate. Various known extensions of the sampling theorem follow as special cases of the resulting generalized sampling expansion of f(t).

C. The Zero-Crossing Problem in Deconvolution[3]

In the determination of the transform $X(\omega)$ of a signal x(t) in terms of the function y(t) and the sum $\underset{\sim}{w}(t) = x(t)*y(t)+\underset{\sim}{n}(t)$, the presence of the term $\underset{\sim}{n}(t)$ (noise) introduces large errors in the vicinity of the zero-crossings ω_i of $Y(\omega)$. A method is presented for reducing these errors. The method is based on the fact that the frequency components $\underset{\sim}{N}(\omega_i)$ of the noise process $\underset{\sim}{n}(t)$ equal the frequency components $\underset{\sim}{W}(\omega_i)$ of the data process $\underset{\sim}{w}(t)$. This leads to an estimate of $\underset{\sim}{N}(\omega)$ for all ω resulting in a reduction of the errors particularly in the vicinity of ω_i.

D. The Factorization Problem for Time-Limited Functions and Trigonometric Polynomials[4]

The following form of the factorization problem is considered: Given a function g(t) vanishing for $|t| > a$ and with non-negative Fourier transform $G(j\omega)$, find a function f(t) with energy spectrum $G(j\omega)$ and such that $f(t) = 0$ outside the interval (0, a). A numerical method for determining f(t) is developed involving only discrete Fourier series.

E. Identification of Systems Driven by Non-Stationary Noise[5]

The system function of an unknown system can be determined in terms of the power spectrum of the output if the input $\underset{\sim}{n}(t)$ is a random process with known properties. This well-known method of system identification is based on the assumption that $\underset{\sim}{n}(t)$ is

stationary. In this paper, it is shown that a system with time-limited impulse response can be identified in terms of the running spectrum of the output even if the input is non-stationary. The analysis includes the evaluation of the autocovariance of the amplitude spectrum and the energy spectrum of a random process. The variance of the estimates is determined under the assumption that the input process is normal. However, the results hold for a more general class of inputs if the time of observation is sufficiently large.

Advanced Research Projects Agency
N00014-76-C-0144

A. Papoulis

REFERENCES

1. A. Papoulis, "The Two-to-One Rule in Data Smoothing," IEEE Trans. Inform. Theory, Vol. IT-23, No. 5 (September 1977).
2. A. Papoulis, "Generalized Sampling Expansion," IEEE Trans. Circuits Systems, Vol. CAS-24, No. 11 (November 1977).
3. A. Papoulis, "The Zero-Crossing Problem in Deconvolution," IEEE Trans. Inform. Theory, Vol. IT-24, No. 1 (January 1978).
4. A. Papoulis, "The Factorization Problem for Time-limited Functions and Trigonometric Polynomials," IEEE Trans. Circuits Systems, Vol. CAS-25, No. 1 (January 1978).
5. A. Papoulis, "Identification of Systems Driven by Non-Stationary Noise," IEEE Trans. Inform. Theory, Vol. IT-24, No. 1 (March 1978).

DIFFRACTION ELIMINATION FOR POINT SOURCES

A. Papoulis and C. Chamzas

A method is developed for recovering the precise location and intensity of point sources in terms of their image greatly distorted by diffraction effects and noise.[1,2] The method is based on an iteration scheme that restores the missing high frequencies of the unknown object.

Consider a one-dimensional optical system with amplitude spread function h(x). With the usual assumptions,[3] the amplitude g(x) of the image of a coherent object is given by

$$g(x) = \int_{-\infty}^{\infty} f(x-\xi)\,h(\xi)\,d\xi \tag{1}$$

where f(x) is the amplitude of the object.

Using capital letters for the Fourier transforms of the above functions, we conclude from Eq. (1) that

$$G(u) = F(u)\,H(u) \tag{2}$$

The case which we are going to examine here is when the object is a double or multiple star and the optical system a telescope. Hence, if

$$d = 2a \tag{3}$$

is the diameter of the lense of the telescope then[3] the Fourier transform of the spread function h(x) is (see Fig. 1)

$$H(u) = \begin{cases} 1 & |u| < a \\ 0 & |u| > a \end{cases} \tag{4}$$

and the amplitude of the object f(x) is given by

$$f(x) = \sum_{i=1}^{K} a_i\,\delta(x - x_i) \tag{5}$$

From the properties of Fourier transform it follows that h(x) is of infinite extent. Thus, a consequence of Eq. (4) is the diffraction effect, yielding as image, g(x), a smooth version of the object.

$$g(x) = \sum_{i=1}^{K} a_i\,h(x - x_i) \tag{6}$$

If x_i are close to each other then the smoothing of the image makes the stars indistinguishable.

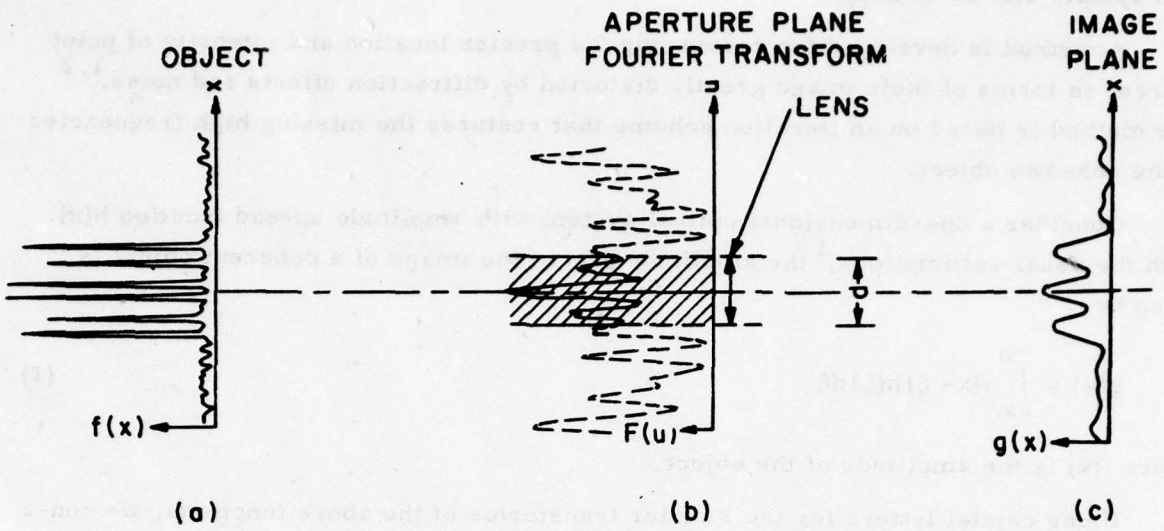


Fig. 1. Schematic of the diffraction problem.
(a) Unknown object f(x).
(b) Fourier transform F(u) of f(x).
(c) Resulting image g(x).
d: Lens diameter.

The developed algorithm using the assumption of Eq. (5) achieves a restoration of the missing high frequencies, for ($|u| > a$), and therefore recovers completely the object. Below we are going to look at the specific case of double stars when their image is greatly distorted by the diffraction effects and noise makes the object appear as a single star (i.e., the resolving power of the telescope can not separate the two stars). The iteration has the following scheme

$$F_n(u) = \begin{cases} F(u)\,H(u) & \\ [1 - H(u)]\,G_n(u) & |u| > a \end{cases} \quad \text{with } 0 < H(u) \leq 1 \tag{7}$$

and

$$g_{n+1}(x) = \begin{cases} f_n(x) & \text{if } f_n(x) \geq C_n \\ 0 & \text{if } f_n(x) < C_n \end{cases} \tag{8}$$

where $f_n(x)$ is the inverse Fourier transform of $F_n(u)$ and C_n an adaptive truncating level, determined by the algorithm. A complete description of the algorithm is given in Reference 2. Hence, in Eq. (7) with the assumption of coherent object

$$H(u) = \begin{cases} 0 & |u| > a \\ 1 & |u| < a \end{cases} \tag{4}$$

If the object is incoherent with illuminance f(x) then,[3,4] the illuminance g(x) of its image is

$$g(x) = \int_{-\infty}^{\infty} f(\xi)\,\bar{h}(x-\xi)\,d\xi \tag{9}$$

with

$$\bar{h}(x) = |h(x)|^2$$

Thus from Fourier transform properties we obtain

$$\bar{H}(u) = \begin{cases} 1 - \dfrac{|u|}{2a} & |u| \le 2a \\ 0 & |u| > 2a \end{cases} \tag{10}$$

and in the relation Eq. (7) of the iteration, in this case, we must use $\bar{H}(u)$ instead of H(u).

The versatility of inserting H(u) inside the algorithm Eq. (7) makes it feasible to skip the deconvolution problem.

Example: The simulated coherent object f(x) is shown in Fig. 2(a) with

$$f(x) = \delta(x - x_i) + 0.5\,\delta(x - x_i - 4\Delta x) + \underset{\sim}{n}(t) \tag{11}$$

where Δx is the sampling interval and $\underset{\sim}{n}(t)$ is white noise uniformly distributed from $[-0.1, 0.1]$. The Fourier transform $F(u) = R(u) + jX(u)$ is shown in Figure 2(b). Figure 3(a) shows the image g(x) of the object and Fig. 3(b) shows

$$G(u) = F(u)\,H(u) = \begin{cases} F(u) & |u| < a \\ 0 & |u| > a \end{cases} \tag{12}$$

the Fourier transform of the image.

In Fig. 4(a), (b) and (c) the 0-th, 4-th and 30-th step of the iteration are shown. The object has been recovered completely at the end of the 30-th iteration.

In the case when the two stars are of the same power $a_1 = a_2$ the limits of the iteration have been calculated numerically. The unknown object consists of two equal pulses located 7 sampling intervals, (Δx), apart and the added noise is white, uniformly distributed. Thus the unknown object f(x) is

$$f(x) = \delta(x - x_1) + \delta(x - x_1 - 7\Delta x) + \underset{\sim}{n}(t) = s(t) + \underset{\sim}{n}(t) \tag{13}$$

In the diagram of Fig. 5 the vertical axis is the S/N ratio and the horizontal axis is analog to the diameter of the lense. A FFT (Fast Fourier Transform) of 256 points

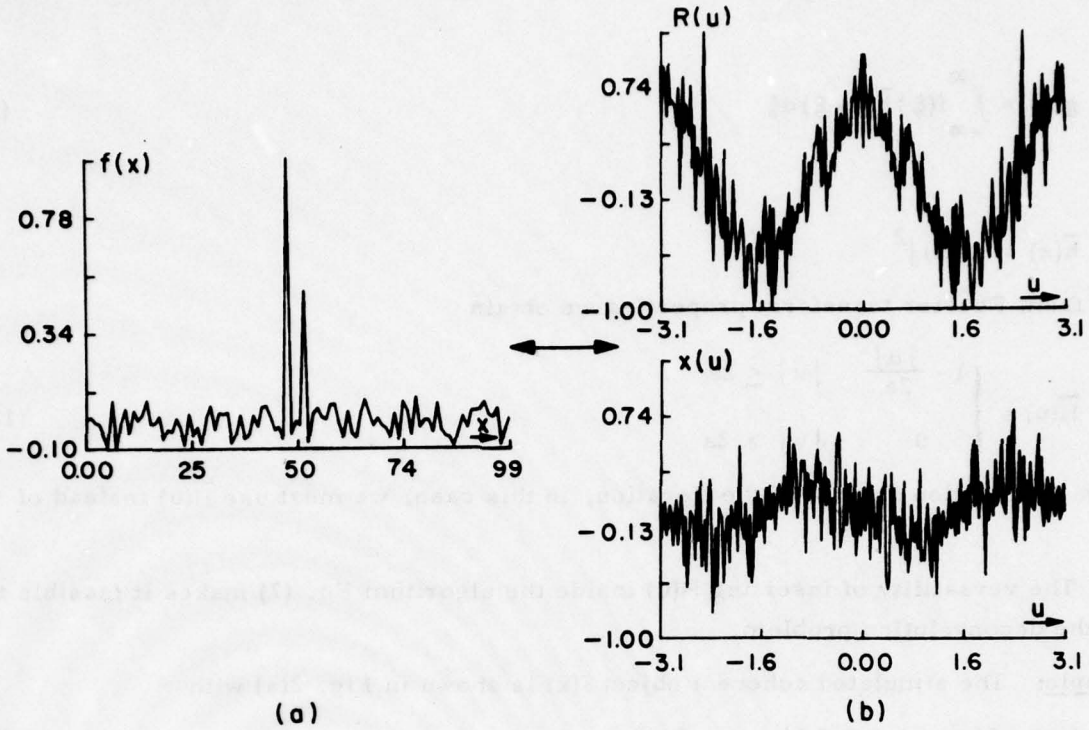


Fig. 2. The unknown object f(x) and its Fourier transform f(u) = R(u) + j X(u).

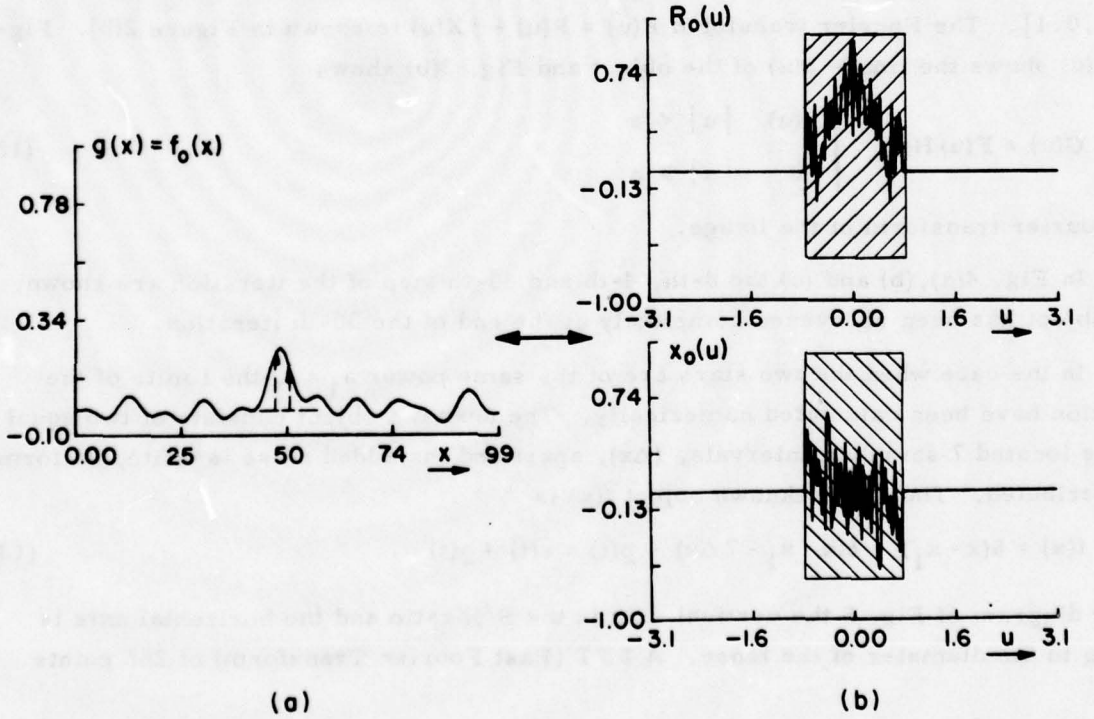


Fig. 3. The observed image g(x) and its Fourier transform $G(u) = R_o(u) + j X_o(u)$.

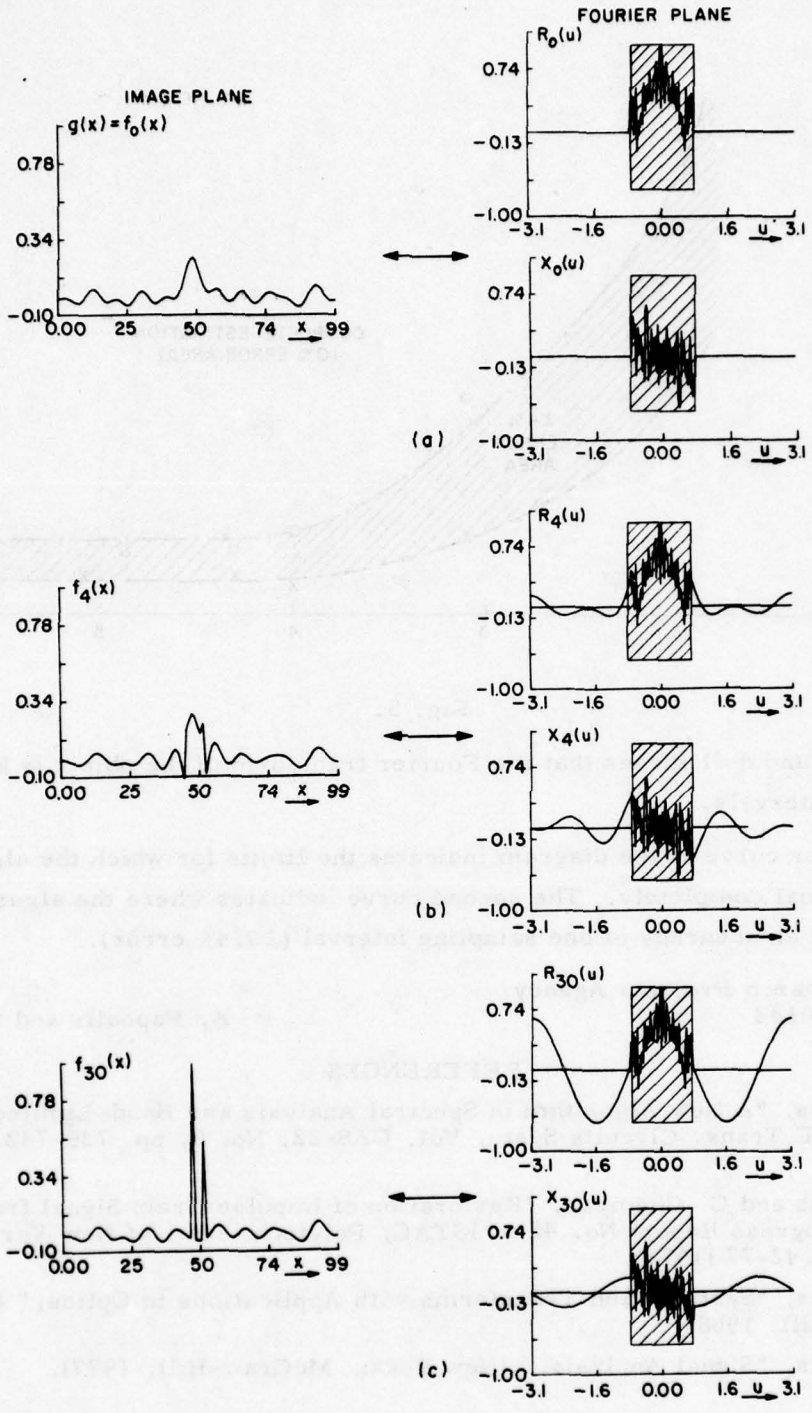


Fig. 4. The iteration.
(a) Given image and its Fourier transform. (Real and Imaginary part.)
(b) The result of the 4-th iteration.
(c) The result of the 30-th iteration.

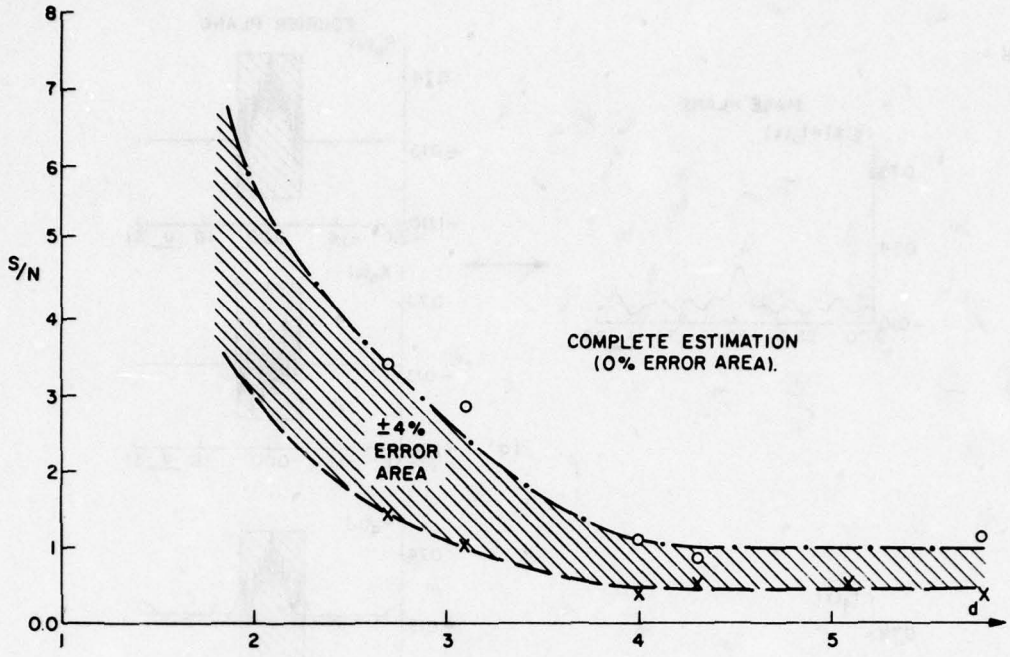


Fig. 5.

has been used and d = 1 means that the Fourier transform of the object is known only for 10 sampling intervals.

The upper curve in the diagram indicates the limits for which the algorithm recovers the signal completely. The second curve indicates where the algorithm recovers the signal with an accuracy of one sampling interval ($\pm 0.4\%$ error).

Advanced Research Projects Agency
N00014-76-C-0144

A. Papoulis and C. Chamzas

REFERENCES

1. A. Papoulis, "A New Algorithm in Spectral Analysis and Band-Limited Extrapolation," IEEE Trans. Circuits Syst., Vol. CAS-22, No. 9, pp. 735-742 (September 1975).

2. A. Papoulis and C. Chamzas, "Restoration of Impulse Train Signal from Smoothed Data," Progress Report No. 42 to JSTAC, Polytech. Inst. of New York, Report No. R-452.42-77 (1977).

3. A. Papoulis, "Systems and Transforms with Applications in Optics," (New York: McGraw-Hill, 1968).

4. A. Papoulis, "Signal Analysis," (New York: McGraw-Hill, 1977).

A CONTRIBUTION TO EDGE DETECTION USING TWO-WAY ANOVA TECHNIQUES

P. Legakis and L. Kurz

In this report, the edge detection techniques using ANOVA techniques in conjunction with quantile statistics, described by the authors previously,[1] are generalized from one-way to two-way ANOVA procedures. The generalization is of importance if the image to be processed is severely corrupted by noise. Since efficient quantile estimation techniques are readily available,[2, 3] and reduction and preconditioning of the data set by quantile partitioning makes the simple fixed effects model of ANOVA applicable to gray level and texture edge detection problems, the procedure described here is particularly attractive in practical applications.

A. Mathematical Formulation of the Problem

Using the same notation and physical modelling as in Ref. 1, the framework for the generalized procedure follows the usual steps in bridging the gap between the one-way and two-way ANOVA procedures.

In the two-way ANOVA case, the point of interest (r, s) is the center of a neighborhood which in this case is a rectangular array of M_A rows and M_B columns. Thus, the rectangular array ${}_{rs}X$ of the neighborhood points consists of elements $\{{}_{rs}X_{ij}\}$ such that

$$
\begin{aligned}
&{}_{rs}X_{ij} = q_{r-w+i;\ s-z+j} \\
&i = 1, 2, \cdots, M_A;\ j = 1, 2, \cdots, M_B \\
&r = w, w+1, \cdots, L_{m-w};\ s = z, z+1, \cdots, L_{n-z} \\
&w = M_A \div 2\ ;\ z = M_B \div 2
\end{aligned}
\tag{1}
$$

Adopting the same notation as for the one-way case.

$$
\begin{aligned}
&{}_{rs}X = q_{r-w+jM_A;\ s-z+iM_B} \\
&r = w, w+1, \cdots, L_{m-w};\ s = z, z+1, \cdots, L_n - z
\end{aligned}
\tag{2}
$$

Here, again, the expression for the quantiles has M_A row indices and M_B column indices. The restriction on the values of r and s follows the same reasoning as for the one-way case and it means that the first and last (w-1) rows and the first and last (z-1) columns will not be classified but they will be used in the classification of points beginning from (w, z) and terminating with the point (L_m - w, L_n - z). The requirements imposed on the scanner resolution (N_r, N_s) are that N_s be much greater than M_B x n and that N_r be much greater than M_A, where n is the data reduction factor. If the above restrictions are

met and if the minimum distance between successive vertical and horizontal edges is much greater than $M_B \times n$ and M_{AS} respectively, the effect of ignoring all border points in the classification as well as false indications of edges due to the presence of actual edges in the neighborhood will be minimized or eliminated.

The corresponding hypothesis testing problem for the two-way ANOVA case take the form:

H_A	rows are homogeneous	→	no horizontal edge
H_B	columns are homogeneous	→	no vertical edge
k_A	rows are heterogeneous	→	horizontal edge
k_B	columns are heterogeneous	→	vertical edge

and the pertinent statistics for the neighborhood points base on (r, s) are

$$rs\overline{X}_{i.} = \frac{1}{M_B} \sum_{j=1}^{M_B} q_{r-w+i;\ s-z+j} \tag{3}$$

$$rs\overline{X}_{.j} = \frac{1}{M_A} \sum_{i=1}^{M_A} q_{r-w+i;\ s-z+j} \tag{4}$$

$$rs\overline{X}_{..} = \frac{1}{M_A M_B} \sum_{i=1}^{M_A} \sum_{j=1}^{M_B} q_{r-w+i;\ s-z+j} \tag{5}$$

$$rsS_A^2 = M_B \sum_{i=1}^{M_A} rs\overline{X}_{i.}^2 - M_A M_B \overline{X}_{..}^2 \tag{6}$$

$$rsS_B^2 = M_A \sum_{j=1}^{M_B} rs\overline{X}_{.j}^2 - M_A M_B \overline{X}_{..}^2 \tag{7}$$

$$rsS_W^2 = \sum_{i=1}^{M_A} \sum_{j=1}^{M_B} (q_{r-w+i;\ s-z+j} - rs\overline{X}_{i.} - rs\overline{X}_{.j} + rs\overline{X}_{..})^2 \tag{8}$$

$$rsf_A = \frac{(M_B - 1)_{rs}S_A^2}{rsS_w^2} \tag{9}$$

$$rsf_B = \frac{(M_A - 1)_{rs}S_B^2}{rsS_w^2} \tag{10}$$

The statistics ${}_{rs}f_A$ and ${}_{rs}f_B$ are sensitive to the presence of edges in the horizontal and diagonal, or vertical and diagonal directions, respectively; namely,

$${}_{rs}f_A < T_A \quad \text{and} \quad {}_{rs}f_B < T_B \rightarrow \text{no edge}$$

$${}_{rs}f_A \geq T_A \quad \text{and} \quad {}_{rs}f_B < T_B \rightarrow \text{horizontal edge}$$

$${}_{rs}f_A \geq T_A \quad \text{and} \quad {}_{rs}f_B \geq T_B \rightarrow \text{diagonal edge}$$

$${}_{rs}f_A < T_A \quad \text{and} \quad {}_{rs}f_B \geq T_B \rightarrow \text{vertical edge}$$

Due to asymptotic normality, the ${}_{rs}f$ statistics have the central and non-central ${}_{rs}F$ distributions with $(M_A - 1)$, $(M_A - 1)(M_B - 1)$ degrees of freedom in the case of ${}_{rs}F_A$ and $(M_B - 1)$, $(M_A - 1)(M_B - 1)$ in the case of ${}_{rs}f_B$. If small sample sizes are used, the asymptotic normality assumptions are no longer valid but the expressions for ${}_{rs}f_A$ and ${}_{rs}f_B$ are still useful as means for obtaining possible edges. The choice of the appropriate thresholds, T_A, T_B as well as the choice of neighborhood size M_A, M_B is best done through simulation of known patterns.

B. Simulation Results

In this section, the advantages of using quantile statistics in two-way ANOVA techniques over simple ANOVA techniques applied to the original data for the solution of gray level and texture edge detection problems are demonstrated. The reference pattern is the same as given in Fig. 1 of Reference 1. If the signal-to-noise ratio is low, the quantile techniques based on two-way ANOVA procedures perform very well (compare Figs. 1 and 2, also Figures 3 and 4). Similar results are obtained for problems involving texture edge detection (see Figures 5 and 6).

In general, the simulation studies point to the fact that quantile-ANOVA techniques are much more efficient than simple ANOVA techniques for the solution of edge detection problems. In severe noise environment the use of two-way quantile-ANOVA techniques is mandated if meaningful edge detection is to be achieved.

Joint Services Technical Advisory Committee
F44620-74-C-0056

P. Legakis and L. Kurz

REFERENCES

1. L. Kurz and P. Legakis, "Edge Detection Using ANOVA Techniques in Conjunction with Quantile Statistics," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76, pp. 328-336 (1976).

2. P. Kersten and L. Kurz, "Robustized Vector Robbins-Monro Algorithm with Applications to m-Interval Detection," Information Sciences, Vol. 11, pp. 121-140 (1976).

3. P. Kersten and L. Kurz, "Bivariate m-Interval Classifiers with Applications to Edge Detection," Information and Control (June 1977).

Fig. 1. Edge detection using two-way ANOVA. Reference pattern corrupted as in Fig. 2 of Reference 1.

Fig. 2. Edge detection using two-way ANOVA after quantile reduction. Reference pattern corrupted as in Fig. 2 of Reference 1.

Fig. 3. Edge detection using two-way ANOVA. Reference pattern corrupted as in Fig. 4 of Reference 1.

Fig. 4. Edge detection using two-way ANOVA after quantile reduction. Reference pattern corrupted as in Fig. 4 of Reference 1.

Fig. 5. Texture edge detection using two-way ANOVA. Reference pattern corrupted as in Fig. 6 of Reference 1.

Fig. 6. Texture edge detection using two-way ANOVA after quantile reduction. Reference pattern corrupted as in Fig. 6 of Reference 1.

M-GRAY LEVEL DETECTION FOR MODERATE AND HIGH SIGNAL-TO-NOISE RATIOS

P. Legakis and L. Kurz

The edge detection and the associated two-gray level detection problems have been treated by the authors elsewhere.[1, 2, 3] The generalization of the edge detection procedures from two- to M-gray levels is straightforward but there is a difficulty associated with this extension: though edges are detected, there is no indication what gray levels are enclosed by the edges. This fact requires further methodological refinements which permit a generalization of the gray level detection problem from $M = 2$ to $M > 2$. In Section A, the noiseless M-gray level detection is reviewed which sets the stage for the treatment of the noisy case. In Section B, the general problem of M-gray level detection under noisy conditions is stated and placed in a mathematical framework of hypothesis testing. In Section C, a simple sequential procedure for M-gray level detection is outlined, a test statistic which yields good performance for moderate and high signal-to-noise is introduced. In Section D, a threshold for the statistic which minimizes the average probability of gray level misclassification is derived. Finally, computer simulations in support of the theory presented in this report are given.

A. M-Gray Level Detection in the Absence of Noise

Let $R(c_0, c_M)$ be the range of the gray scale and let $I(i,j) \in R(c_0, c_M)$ be the brightness of an image array H at the point (i,j). $I(i,j)$ can be represented by an ℓ-bit quantization number $(b_1, b_2, \cdots, b_\ell)$ where b are binary numbers and $1 \leq \ell \leq 8$. The lower and upper bounds for ℓ were chosen to be 1 and 8, respectively, because a gray scale cannot have less than 2 levels, and $2^8 = 256$ is more than most images require for excellent quality reproduction. Each quantization number could, in the limit, correspond to one gray level. However, since for $\ell = 8$ it would be very difficult to establish and reproduce 256 brightness levels, a reduced scale consisting of $m < 256$ brightness levels is desirable. Thus, a reduced M-gray level scale consists of levels g_k such that

$$c_{k-1} \leq g_k < c_k \quad , \quad k = 1, 2, \cdots, M$$

where

c_0 : quantization number corresponding to the darkest point in the gray scale

c_M: quantization number corresponding to the brightest point in the gray scale.

Denoting by $[g_k]$ the brightness of level g_k the relationship

$$[g_1] < [g_2] < [g_3] < \cdots < [g_M]$$

holds, and for every g_k there is at least one quantization number.

In the absence of noise, the transformation from quantization numbers to gray levels is based on a sample of a single observation and I(i, j), the brightness of an image array at the point (i, j), maps onto level g_k if

$$I(i,j) \in [c_{k-1}, c_k], \quad k = 1, 2, \cdots, M \tag{1}$$

Depending on the gray scale partition, i.e., the choice of $c_1, c_2, c_3, \cdots, c_M$ which involves not only the choice of M but also the actual g_k one could generate different reproductions of the same image. In fact, even if M is fixed and the choice of $c_1, c_2, \cdots, c_M$ made, one could still generate different representations of the same original image array based on the choice of characters that are used in an overstrike arrangement to produce a given brightness level.

If no additional a priori information about the image is available, the assumption is made that the gray scale is partitioned into M equal levels, i.e., into levels which consist of the same number of quantization numbers. This assumption seldom affects the quality of the reproduced image in the noiseless case. As it will be shown in subsequent sections, this assumption does not affect the gray level detection in the noisy case. Logically this insensitivity of gray level detection to the partition of the scale follows from the fact that in an unequal partition of the gray scale, the different levels still differ by a shift-in-the-mean in addition to the change-of-scale. The statistic to be used for any gray level detection is sensitive to both, and unequal partition does not influence the decision process. For simplicity, equal partitioning of the gray scale will be assumed for the remainder of the report. Thus, in the noiseless case, once the gray scale has been partitioned, the identification of an image array point with a particular gray level follows directly from Equation (1). The problem to be resolved is what happens when data is corrupted by noise.

B. The M-Gray Level Classification Problem Under Noise Conditions

The two-gray level classification problem under noise conditions has received considerable attention, and procedures such as analysis of variance,[4] slippage algorithms,[5] as well as the quantile procedures of Refs. 1, 2, 3 are quite adequate for almost any noise environment. The problem was solved by posing it in a hypothesis testing framework with level 0 represented by a distribution of given mean and variance and level 1 by another distribution of different mean and the same variance. Note that the assumption of equal variance amounts to equal partition of the gray scale.

One can generalize the above problem to the M-gray level case by considering a gray scale that is spanned by M distribution functions having different means and the same variances if equal partition is assumed, or having different means and different variances, in the more general case. A mathematical statement of the problem now follows.

Let the gray scale be spanned by M distribution functions $F_1(x), \cdots, F_M(x)$ representing the M-gray levels, and let $F_k(x) = F(x - \mu_k)$ where μ_k is the mean for level k and

$$\mu_1 < \mu_2 < \mu_3 < \cdots < \mu_M$$

or the distributions are all identical except for a shift-in-the-mean. Assume the availability of reference sample X_k and that $X_k \sim F(x - \mu_k)$, $k = 1, 2, \cdots, M$. Let $Y_n = y_1, y_2, y_3, \cdots, y_M$ be a sample from an image array, H, that needs to be classified as coming from one of the $F(x - \mu_k)$. A classification procedure could be either sequential, consisting of at most M steps, or parallel requiring that all M steps be executed simultaneously, or some combination of these procedures. The choice between a sequential and a parallel procedure involves a tradeoff between the higher cost of implementing the parallel procedure and better performance in terms of less waiting time resulting from the simultaneous execution of M steps.

If a sequential procedure is followed, the sample to be classified is compared to the reference sample X_k and a threshold statistic T_k results. Then hypothesis H_k is accepted or rejected on the basis of whether T_k exceeds or does not exceed a threshold. If H_k is accepted, meaning that $X_k \sim F(x - \mu_k)$, the classification stops; otherwise, the test sample is matched with reference sample X_{k+1} until a match occurs or the reference samples are exhausted. In the latter case, where all reference samples have been exhausted without making a classification, a decision is made on the basis of min or max $(T_1, T_2, \cdots, T_M)$. If a parallel procedure is followed, a set of test statistics $T_1, T_2, T_3, \cdots, T_M$ is generated and the test sample is classified on the basis of min or max $(T_1, T_2, \cdots, T_M)$. It must be noted that the alternative to hypothesis H_k is taken to be H_{k+1} not because the gray levels are ordered in real images (they are not) but because only adjacent distributions are expected to cause a misclassification on the hypothesis. Also, H_{k-1} is not considered as an alternative at the kth step because its rejection at the previous stage is the reason why the procedure has reached the kth step.

In the remainder of the report, sequential procedures are explored in detail because they are simpler and more economical to implement. The extension of these to parallel operations follows directly. Usually, sequential procedures are used to minimize the number of samples required before making a decision. Here, all samples are available and the emphasis is on minimizing the number of steps required to arrive at a classification. If, in addition, use is made of the classification of the previous sample as a starting point of comparison, the number of steps required for classification is further reduced because it is expected that neighboring points frequently belong to the same class.

C. M-Gray Level Sequential Pattern Detection

Let $X_k \sim F(x - \mu_k)$ and Y_n be the reference and test samples, respectively, consisting of n.i.i.d.r.v. observations each. The test statistic, T_k, for the kth step of the procedure is defined as follows:

On the basis of X_k, let $[q_{k,50-\gamma}, q_{k,50+\gamma}]$ be a quantile interval centered* around the 50th percentile, where

$$\int_{-\infty}^{q_{k,50-\gamma}} dF(x - \mu_k) = p_{k,50-\gamma}$$

Sample Y_n consists of points from at most two distributions corresponding to any two gray levels. The cases of interest are those for which the quantile interval contains observations from adjacent distributions. In all other cases the contributions to misclassification are considered negligible. The latter case corresponds, roughly, to high signal-to-noise ratio conditions.

The test statistic T_k is now formed by counting the number of observations in Y_n that fall inside $[q_{k,50-\gamma}, q_{k,50+\gamma}]$ i.e.,

$$T_k = \text{Card}\{y_j: \; q_{k,50-\gamma} \leq y_j < q_{k,50+\gamma}\}$$

$$k = 1, 2, \cdots, M; \quad j = 1, 2, \cdots, n$$

The formulation of the test statistic is distribution-free since quantile information rather than exact knowledge of the underlying distribution is required, from an implementation point of view, T_k is very simple: it requires operations of the order of O(n) once the quantile interval is established. A simple procedure for establishing the quantiles from the observed data even if the data is contaminated is available in Ref. 6 where dependence in m-interval detectors is also treated.

Under the hypothesis (H_k), the observations of X_k and Y_n are generated by the same continuous distribution $F(x - \mu_k)$. From the classical "occupancy" model which is extensively treated in the literature, the statistic T_k is binomially distributed, i.e., $T_k \sim B(n, p_{H_k})$ with p_{H_k} denoting the probability that an observation from H_k falls within the prespecified quantile interval $[q_{k,50-\gamma}, q_{k,50+\gamma}]$ and n is the total number of trials. Thus,

---

*The 50th percentile is chosen because that is where most of the information is concentrated in the shift-in-the-mean problem.

$$\text{Prob}_{H_k}(T_k) = \begin{cases} \binom{n}{T_k} p_{H_k}^{T_k} q_{H_k}^{n-T_k} & T_k = 0, 1, \cdots, n \\ 0 & \text{Oth.} \end{cases}$$

and

$$q_{H_k} = 1 - p_{H_k}$$

For example, if $n = 20$ and $p_{H_k} = q_{H_k} = .5$, $\text{Prob}_{H_k}(0) = 2^{-20}$ which is very small, as expected.

The reason why the classification procedure has reached the kth step is that H_k, $k = 1, 2, \cdots, k-1$ have been rejected. The alternative at the kth step is H_{k+1} and the distribution of T_k under H_{k+1} is assumed to also be of the binomial form, i.e.,

$$T_k \sim B(n, p_{H_{k+1}})$$

where

$$p_{H_{k+1}} = \theta_{k+1} p_{H_k}$$

and

$$\theta_{k+1} \quad \frac{\dfrac{1}{|\mu_{k+1} - \mu_k| + G_{H_{k+1}}}}{G_{H_k}}$$

The formulation of the distribution of T_k under the alternative is also in agreement with what is intuitively expected. Namely, for large signal-to-noise ratios, θ_{k+1} is small and so is $p_{H_{k+1}}$. When the alternative approaches the hypothesis, $\theta_{k+1} \to 1$ and $T_{k/H_{k+1}} \to T_{k/H_k}$. Thus, under the alternative

$$\text{Prob}_{H_{k+1}}(T_k) = \begin{cases} \binom{n}{T_k} p_{H_{k+1}}^{T_k} q_{H_{k+1}}^{n-T_k} & T_k = 0, 1, \cdots, n \\ 0 & \text{Oth.} \end{cases}$$

D. Classification Procedure for the kth Step and a Threshold Specification

As it was noted in the previous section, the classification procedure has reached the kth step because the first (k-1) alternatives have been rejected. The classification for the kth step proceeds as follows:

(1) From Y_n and $[q_{k,50-\gamma}, q_{k,50+\gamma}]$, compute T_k

(2) Compare T_k with T_{ko}

(3) Decide that H_k is true if $T_k \geq T_{ko}$ or that H_{k+1} is true (only temporarily until verified in the subsequent step) if $T_k < T_{ko}$.

The threshold T_{ko} is chosen to minimize the average probability of misclassification

$$P_{ek} = \text{Prob}(H_k)\,\text{Prob}(H_{k+1}/H_k) + \text{Prob}(H_{k+1})\,\text{Prob}(H_k/H_{k+1}) \tag{2}$$

where

$\text{Prob}(H_k)$: a priori probability of H_k

$\text{Prob}(H_{k+1}/H_k)$: type I error

$\text{Prob}(H_{k+1})$: a priori probability of H_{k+1}

$\text{Prob}(H_k/H_{k+1})$: type II error

It is customary to assume $\text{Prob}(H_k) = \frac{1}{M}$, $\forall k$, unless additional information is available. Since

$$\alpha_k = \text{Prob}(H_{k+1}/H_k) = \sum_{T_k=0}^{T_{ko}} \text{Prob}_{H_k}(T_k)$$

and

$$\beta_k = \text{Prob}(H_k/H_{k+1}) = \sum_{T_{ko}+1}^{n} \text{Prob}_{H_{k+1}}(T_k)$$

Equation (2) becomes

$$P_{ek} = \frac{1}{M}\left(\sum_{T_k=0}^{T_{ko}} \text{Prob}_{H_k}(T_k) + \sum_{T_{ko}+1}^{n} \text{Prob}_{H_{k+1}}(T_k)\right)$$

$$= \frac{1}{M}\left(\sum_{T_k=0}^{T_{ko}} \text{Prob}_{H_k}(T_k) + 1 - \sum_{T_k=0}^{T_{ko}} \text{Prob}_{H_k}(T_k)\right) \tag{3}$$

where

$$\text{Prob}_{H_k}(T_k) = \begin{cases} \binom{n}{T_k} p_{H_k}^{T_k} q_{H_k}^{n-T_k} & T_k = 0, 1, 2, \cdots, n \\ 0 & \text{Oth.} \end{cases}$$

$$\text{Prob}_{H_{k+1}}(T_k) = \begin{cases} \binom{n}{T_k} p_{H_{k+1}}^{T_k} q_{H_{k+1}}^{n-T_k} & \\ 0 & \text{Oth.} \end{cases}$$

It is well known from Bayesian theory that T_{ko} is the solution of

$$\text{Prob}(H_k)\,\text{Prob}(T_k/H_k) = \text{Prob}(H_{k+1})\,\text{Prob}(T_k/H_{k+1})$$

Therefore

$$\frac{1}{M}\binom{n}{T_k}(p_{H_k})^{T_k}(1-p_{H_k})^{n-T_k} = \frac{1}{M}\binom{n}{T_k}(\theta_{k+1}p_{H_k})^{T_k}(1-\theta_{k+1}p_{H_k})^{n-T_k}$$

or

$$\left(\frac{p_{H_k}}{\theta_{k+1}p_{H_k}}\right)\left(\frac{1-\theta_{k+1}p_{H_k}}{1-p_{H_k}}\right)^{T_k} = \left(\frac{1-\theta_{k+1}p_{H_k}}{1-p_{H_k}}\right)^{n} \tag{4}$$

Taking the natural logarithm of both sides of Eq. (4), results in

$$T_k \ell n\left(\frac{p_{H_k}}{\theta_{k+1}p_{H_k}}\,\frac{1-\theta_{k+1}p_{H_k}}{1-p_{H_k}}\right) = n\ell n\left(\frac{1-\theta_{k+1}p_{H_k}}{1-p_{H_k}}\right) \tag{5}$$

or

$$T_{ko} = \frac{n\ell n\left(\dfrac{1-\theta_{k+1}p_{H_k}}{1-p_{H_k}}\right)}{\ell n\left(\dfrac{1}{\theta_{k+1}}\,\dfrac{1-\theta_{k+1}p_{H_k}}{1-p_{H_k}}\right)} \tag{6}$$

Equation (6) gives the threshold T_{ko} which minimizes the average probability of misclassification in the kth step. Since T_{ko} as given above is not likely to be an integer and since it must be constrained to integer values, T_{ko} is given in terms of $\lceil T_{ko} \rceil$ and

$\lfloor T_{ko} \rfloor$ where

$\lfloor T_{ko} \rfloor$ is the largest integer less than T_{ko}

$\lceil T_{ko} \rceil$ is the largest integer greater than T_{ko}.

To select between $\lfloor T_{ko} \rfloor$ and $\lceil T_{ko} \rceil$ one has to substitute into Eq. (3) and retain the value which minimizes it. The parameter θ_{k+1} ranges from 0 to 1. If $\theta_{k+1} = 1$, Eq. (6) takes an indeterminate form, while for $\theta_{k+1} = 0$, $T_{ko} = 0$ as expected. The total probability of error misclassification is minimized when it is minimized in each of the k steps, and

$$P_e \cong \frac{1}{M} \left(\sum_{k=1}^{M} (P_{ek})_{min} \right) \tag{7}$$

where $(P_{ek})_{min}$ is given by Eq. (3) with the threshold specified by Equation (6).

Figure 2 demonstrates the performance of the procedure described in this report under moderate noise conditions. The reference pattern is given in Figure 1.

Joint Services Technical Advisory Committee
F44620-74-C-0056

P. Legakis and L. Kurz

IBM

REFERENCES

1. L. Kurz and P. Legakis, "Edge Detection Using ANOVA Techniques in Conjunction with Quantile Statistics," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76, pp. 328-336 (1976).
2. L. Kurz and P. Legakis, "Gray Level Detection Using M-Interval Occupancy Vector in Conjunction with Mann-Whitney Statistic," Progress Report No. 41 to JSTAC, Polytech. Inst. of New York, Report No. R-452.41-76, pp. 337-347 (1976).
3. L. Kurz and P. Legakis, "A Contribution to Edge Detection Using Two-Way ANOVA Techniques," Progress Report No. 42 to JSTAC, Polytech. Inst. of New York, Report No. 452.42-77 (1977).
4. J. Aron and L. Kurz, "A Statistical Approach to Edge Detection and Enhancement," 1973 International Symposium on Information Theory, Ashkelon, Israel (June 1973).
5. R.J. Jakubek and L. Kurz, "Fixed and Variable Threshold Slippage Procedures for Block Pattern Classification and Edge Detection," 1973 International Symposium on Information Theory, Ashkelon, Israel (June 1973).
6. P. Kersten and L. Kurz, "Robustized Vector Robbins-Monro Algorithm with Application to m-Interval Detection," Information Sciences, Vol. 11, pp. 121-140 (1976).

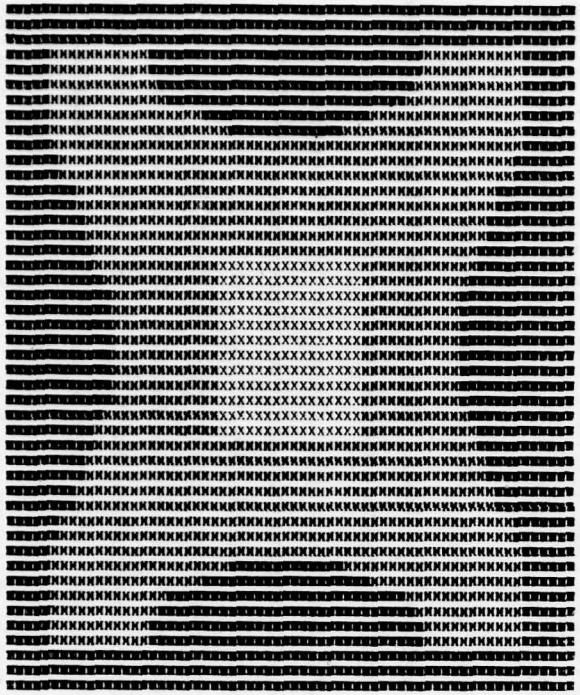


Fig. 1. 3-Gray level reference pattern.

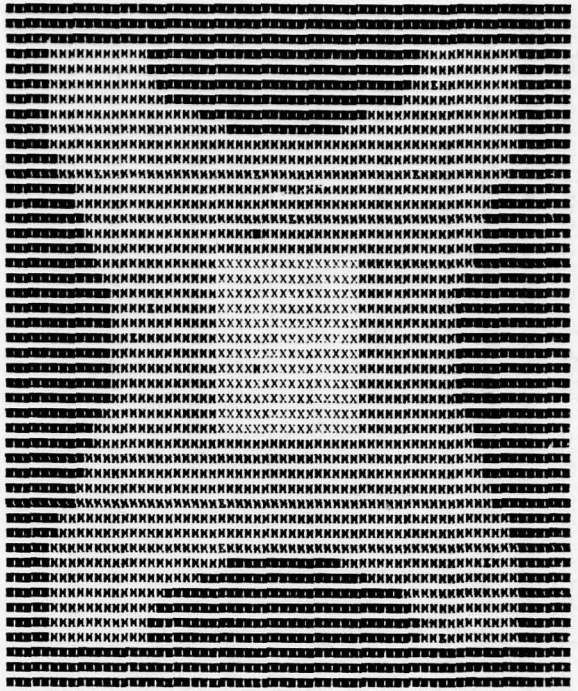


Fig. 2. Gray level detection based on single quantile interval. Reference pattern corrupted with Gaussian N(0,.5), N(1,.5) and N(2,.5) Noise.

BIVARIATE M-INTERVAL CLASSIFIERS WITH APPLICATION TO EDGE DETECTION

P. Kersten and L. Kurz

In 1972 Ching and Kurz[1] introduced a statistic which is a generalization of the sign test. The sign test, one of the first nonparametric statistics, is usually applied when testing the hypothesis $M=M_o$ against the alternative $M > M_o$, or $M < M_o$, or $M \neq M_o$, where M_o is the median. Accordingly, the sign test is defined by $\sum_{i=1}^{n} \text{sgn}(X_i - M_o)$ where $\text{sgn}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$ and X_i are independent and identically distributed samples from the population under test. To generalize the sign test one assumes that a priori knowledge of (m-1) quantiles $\{a_1, \ldots, a_{m-1}\}$ such that $F^{-1}(i/m)=a_i$, where F is the cumulative distribution function (CDF) of the sampled distribution, are given. Using these quantiles, the observation space is partitioned into m disjoint regions $A_i = (a_{i-1}, a_i]$, $i=1, \ldots, m$ where $a_o = -\infty = -a_m$. Assuming that F is strictly increasing so that the a_i are unique and the probability density function $f=F'$ is continuous implies $f(a_o)=f(a_m)=0$. With these definitions and hypotheses one defines the m-Interval Partition Statistic as

$$\tau = \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{m} b_i I_{\{X_k \in A_i\}}$$

where b_i are the scores associated with each partition and $I_{\{X_k \in A_i\}}$ is the indicator function of the event $(X_k \in A_i\}$, i.e. $I_{\{X_k \in A_i\}}$ is 1 if $X_k \in A_i$ and 0 otherwise. Now τ may be rewritten as $\tau = \frac{1}{n} \sum_{i=1}^{m} n_i b_i$ where n_i is defined as the number of samples falling into the partition A_i. If $m=2$ and $b_1=-1=-b_2$, then τ is equivalent to the sign statistic provided $E(X_i) = 0$ and $f(x)=f(-x)$ which implies tha $M_o=E(X_i)$.

Ching and Kurz[1] demonstrate that this statistic is nonparametric in the sense that its Type I Error is independent of the functional form of the underlying distribution and robust, i.e., its performance is insensitive to changes in the underlying noise distribution. It is interesting to note that the generalized sign-test classifiers have been treated in a unified framework in Cochrane and Kurz.[2] This framework is general enough to include not only the m-Interval Partition Tests, but also, the linear rank statistics. Both papers discuss the robustness of the m-Interval Partition Test and its efficiency for large sample sizes. Application of rank tests to classification problems have also been considered in Woinsky and Kurz,[3,4] and Chadwick and Kurz.[5]

Hodges[6] obtained a bivariate extension of the sign test and derived its null distribution. The general setting for the application of this statistic is a bivariate test of

hypothesis. Specifically, given two sets of independent bivariate samples, the first from a bivariate distribution function with mean (0, 0) and the second from a bivariate distribution function whose mean may be (0, 0) and (μ, ξ), one seeks to determine from which population the last sample originated. Hodges considers the comparison of two treatments where one has two sets of bivariate samples (x_i', y_i') and (x_i'', y_i''), $i=1,\ldots,n$. From these samples, a single sequence of bivariate samples is formed $\underline{z}_i = (x_i, y_i) = (x_i'' - x_i', y_i'' - y_i')$ which belong to one of the following hypotheses:

H_o : $\underline{z}_i = (x_i, y_i)$ have mean (0, 0), $i=1,\ldots,n$,

H_1 : $\underline{z}_i = (x_i, y_i)$ have mean $(\mu, \xi) \neq (0, 0)$, $i=1,\ldots,n$.

The central idea in this test consists of projecting the bivariate samples onto an axis passing through the origin at an angle θ (hereafter referred to as the θ-axis) and then applying the univariate sign test along this axis for all possible values of $\theta \in [0, 2\pi]$. One then uses the maximum value assumed by the sign test for all θ to accept or reject the hypothesis. If one defines $\underline{1}_\theta$ as the unit vector of an axis at an angle θ with respect to the x-axis, then one can express the Bivariate Sign Test as

$$B = \sup_{0 \leq \theta \leq 2\pi} \sum_{j=1}^{n} \operatorname{sgn}(\underline{z}_j \cdot \underline{1}_\theta)$$

where $\underline{z}_j \cdot \underline{1}_\theta$ is the dot product of the two-dimensional vector $\underline{z}_j$ with $\underline{1}_\theta$, i.e. $\underline{z}_j \cdot \underline{1}_\theta$, is the projection of $\underline{z}_j$ onto the θ-axis. Intuitively, the angle θ at which B is maximum should be approximately $\theta = \tan^{-1}(\xi/\mu)$ if H_1 is true since the shift of mean occurs along this axis. Under the hypothesis, no direction of θ should be favored and thus one expects the supremum to occur at any angle. Unfortunately, both the univariate and bivariate sign tests discard all the magnitude information in the samples. The univariate m-Interval Partition Statistic, retains enough magnitude information to obtain a more efficient test statistic; but, at the same time, does not surrender its robustness since the magnitude information is quantized via the partitioning of the observation space. The m-Interval Partition Test may be extended to the bivariate case in precisely the same manner as the sign test is extended to the bivariate case.

The motivation for considering the tests described above arises from its application to edge detection problems where robustness is an essential ingredient. For instance, in picture processing, the insensitivity of the classifier's performance to changes in the underlying noise distribution is more important than using an optimum test under strict assumptions on the noise distribution function to achieve the most

efficient operation. In Section A, where a more detailed description of the Bivariate m-Interval Partition Statistic is presented, the nature of the test will guarantee its robustness. In Section B, two sets of scores for this classifier are discussed and the performance of this statistic with these scores is the subject of Section C. An application of the Bivariate m-Interval Partition Statistic to the edge detection of a block pattern is considered in Section D along with simulations of the test in various noise environments.

A. Description of the Bivariate m-Interval Partition Statistic

The key to Hodges extension of the sign test to the bivariate case was to project the samples upon the θ-axis and then applying the sign test along this axis for each θ. The supremum of these test statistics is used to accept or reject the hypothesis. Under suitable conditions the m-Interval Partition Statistic is extended in the same manner. Define $A_i(\theta) = (a_{i-1}(\theta), a_i(\theta)]$, $i=1,\ldots,m$ as the partition along the θ-axis so that $F_\theta^{-1}(i/m) = a_i(\theta)$ where F_θ is the marginal distribution of the samples projected upon the θ-axis. One may define

$$T(\theta) = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{m} b_i I\{\underline{Z}_j \cdot \underline{1}_\theta \in A_i(\theta)\}$$

and T' as $\sup_{0 \le \theta \le 2\pi} T(\theta)$.

Throughout this analysis it is assumed that the joint PDF is symmetric in that $f(x,y)=f(-x,-y)$ and that m, the number of partitions is even. The scores b_i are assumed to be odd with respect to the origin. As presently defined, this statistic is not feasible to calculate since $T(\theta)$ must be evaluated for each $\theta \in [0, 2\pi]$ and accordingly m-1 quantiles $a_i(\theta)$ need to be known for each θ. Therefore, one defines $\theta_i = 2\pi(i-1)/k$ and $T = \max_{1 \le i \le k} (T_i)$ where $T_i = T(\theta_i)$. To calculate T only $(m-1)k$ quantiles need to be known. The Bivariate m-Interval Partition Statistic is defined by $\max(T_1,\ldots,T_k)$ which is less than or equal to $\sup_{0 \le \theta \le 2\pi} T(\theta)$, since the supremum is over a larger class. Thus T will not reduce to the Bivariate Sign Test for $b_1 = -1 = -b_2$ and $F^{-1}(1/2) = a_1(\theta) = 0$. However, this presents no problem since T may be considered upon its own merits, using the Bivariate Sign Test only as a bench mark. For computations used in this paper, $|\theta_{i+1} - \theta_i| = \pi/12$ was found to yield acceptable Type I error and power.

For the set of quantiles $a_i(\theta) = F_\theta^{-1}(i/m)$, $i=1,\ldots,m-1$, and $-a_o = \infty = a_m$, one needs to know the loci of these a_i for all θ. Since the joint PDF is symmetric and assuming the joint CDF $F(x,y) = F(x)\,F(y)$, the quantiles may be represented concisely

as in Fig. 1 where $F(\cdot)$ is a Cauchy Distribution. Figure 1 is representative of the patterns one obtains for the quantile loci for other distributions such as the Bivariate Double-Exponential and Gaussian. Space considerations prevent the illustration of the loci for these latter distributions. The intersection of these curves with any axis passed through the origin at angle θ yields the octiles of the marginal CDF $F_\theta(x)$.

Figure 1 illustrates the futility of trying to obtain a small sample distribution in a manner paralleling Hodges[6] or Ching and Kurz.[1] Hodges and later Klotz[7]

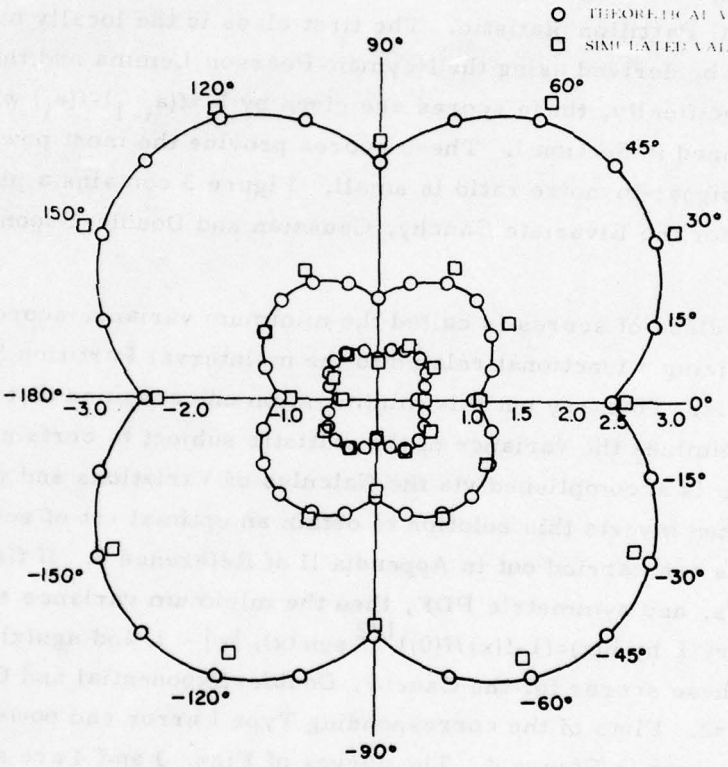


Fig. 1 Plot of quantiles bivariate cauchy-independent components

obtained the small sample distribution by setting up a correspondence between the possible values of $T(\theta)$ and the sample paths of a random walk. Unfortunately, the fact that θ is continuous and that the shape of these regions described is a function of $F(x, y)$ precludes any possibility of establishing such a correspondence for the Bivariate m-Interval Partition Test. The same factors plus the dependence of the $T(\theta_i)$ also forestall the application of generating functions. Thus the small distribution is an open problem.

Practically, the quantiles will not be known a priori. To estimate these

quantiles a multivariate Robbins-Monro procedure is used on a sequence of bivariate samples.[8] This is in essence a training sequence for the statistic which can be updated periodically in real-time systems such as picture transmission by sending a pre-assigned test pattern. Thus the partial characterization of the bivariate distribution via these k(m-1) quantiles is realistic.

B. The Scores

Two classes of scores were investigated for use with the Bivariate m-Interval Partition Test. The investigation of these scores is done in the one-dimensional case with the m-Interval Partition Statistic. The first class is the locally most powerful scores which may be derived using the Neyman-Pearson Lemma and the small signal assumptions. Specifically, these scores are given by $b_i=f(a_{i-1})-f(a_i)$ where all these quantities are defined in Section I. These scores provide the most powerful test of hypothesis if the signal-to-noise ratio is small. Figure 3 contains a plot of the Type I Error and power for the Bivariate Cauchy, Gaussian and Double-Exponential distribution functions.

The second class of scores is called the minimum variance scores. These are derived by minimizing a functional related to the m-Interval Partition Statistic at or near the hypothesis. To carry out this minimization one assumes that the scores are given and minimizes the variance of the statistic subject to certain constraints. This minimization is accomplished via the Calculus of Variations and yields the optimal PDF. One then inverts this solution to obtain an optimal set of scores given this PDF. The details are carried out in Appendix II of Reference 8. If f(x) is a unimodal, continuous, and symmetric PDF, then the minimum variance scores are given for arbitrary L by $b(x)=(1-f(x)/f(0))^{1/2}\,\mathrm{sgn}(x)$, $|x| < L$ and $\mathrm{sgn}(x)$ for $|x| > L$. Figure 2 shows these scores for the Cauchy, Double-Exponential and Gaussian distributions with L=2. Plots of the corresponding Type I error and power for the same distributions are given in Figure 4. The curves of Figs. 3 and 4 are simulated and their generation is discussed in Section V. In application, b(x) is evaluated at the quantiles in order to determine b_i.

Observe that the Type I Error or α-curve varies for different distributions and thus by the definition given in Section I, the Bivariate m-Interval Partition Test is not nonparametric in the sense of fixed Type I Error. However, the one-dimensional m-Interval Partition Test is nonparametric (Ching and Kurz, 1972). This property is lost in the bivariate extension due to the dependence of $T(\theta)$ on θ. However, it is the partitioning of the θ_i-axes and the definition of the statistic as a set function on these partitions that gives T its robustness. This important property is retained.

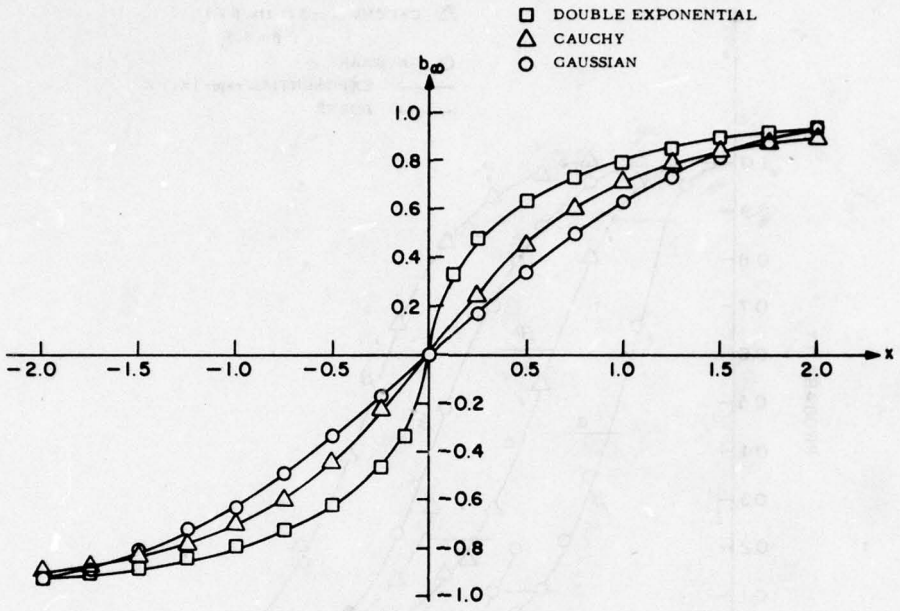


Fig. 2 Minimum variance scores

This fact is reflected in the plot of the minimum variance scores contained in Fig. 2 in that the scores appear relatively insensitive to changes in the distribution.

C. Estimating the Threshold and Probability of Error for the Bivariate m-Interval Partition Statistic

Since the small sample distribution of the Bivariate m-Interval Partition Statistic is unknown, one cannot obtain a set of theoretical curves of the Type I Error denoted by α and the power denoted by β versus the threshold. However, by simulation, the α and β curves may be obtained by the Robbins-Monro procedure using the regression function $M(\alpha) = P(\max_j T_j > \nu \,|H_o) - \alpha$ where T_j is the m-Interval Partition Statistic evaluated along the θ_j- axis and ν is the threshold. The resulting recursion is

$$\alpha_{n+1} = \alpha_n + \frac{1}{n}\left[I_{(\max_j T_j > \nu)} - \alpha_n\right]$$

under H_o. A similar recursion equation holds under H_1 for the derivation of the β-curves. Figure 3 gives the plots of the Type I error and power versus the threshold for the Bivariate Cauchy, Gaussian and Double-Exponential distribution using the locally most powerful scores. Figure 4 contains similar plots for the minimum variance scores. Table I contains a list of the Type II errors and total probability of

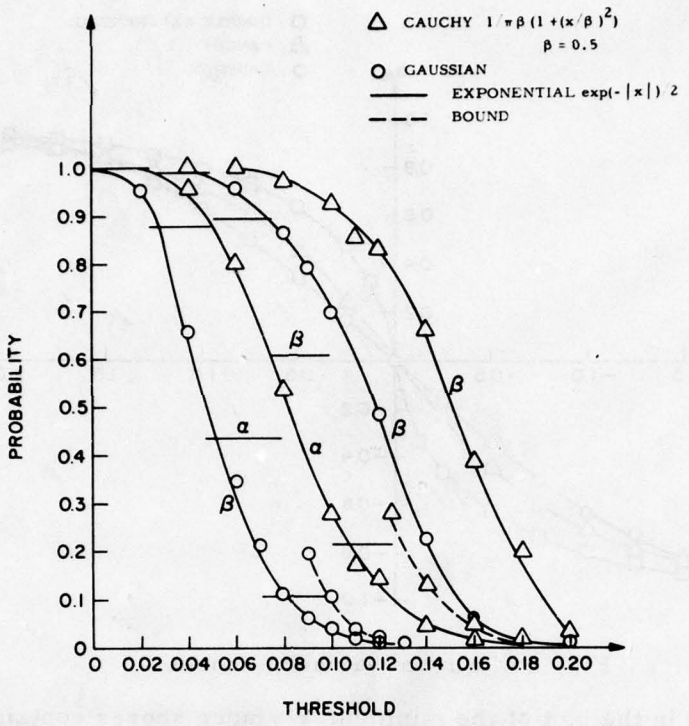


Fig. 3 Plot of type I error and power vs threshold for the locally most powerful scores N=500 iterations

TABLE I

Distribution	Probability of error		Total Probability of error	$(\mu^2+\xi^2)/\sigma_1^2$
	Type I	Type II		Signal to noise ratio
Gaussian	0.11	0.3	0.205	1.0
Cauchy Scale 0.5	0.10	0.22	0.160	undefined
Exponential	0.105	0.39	0.248	0.5

error assuming equally probable hypotheses for approximately equal Type I error. This table was derived directly from Fig. 4 for a sample size of n=10. Note particularly the low signal-to-noise ratios when examining this table and Figures 3 and 4. These results represent the performance of the Bivariate m–interval Partition Classifier in severe noise environments.

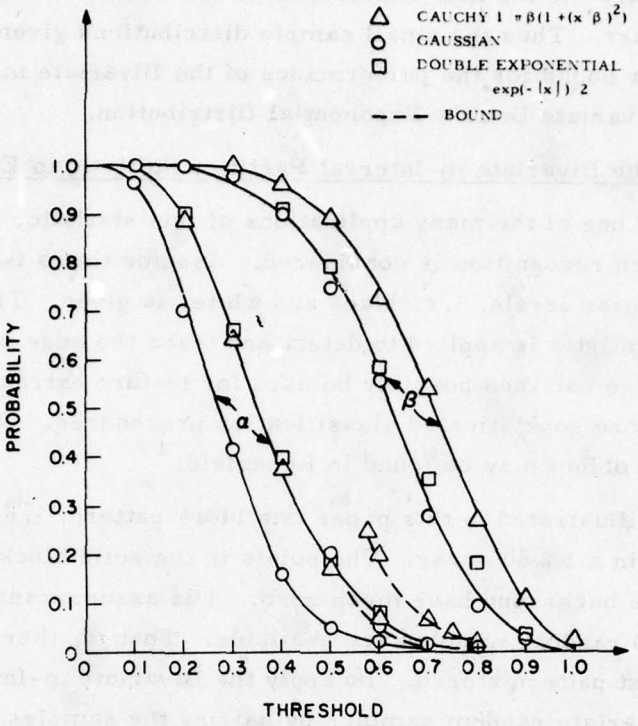


Fig. 4 Plot of type I error and power vs threshold for the minimum variance scores N=500 iterations

One can obtain a bound of the Type I Error by using a result proven by Andre Tchen.[9] This result gives a bound on the probability that a maximum of a sequence of dependent or independent random variables exceeds a threshold ν. Let $M_n = \max(X_1, \ldots, X_n)$, then $P(M_n > \nu) = P(\bigcup_{i=1}^{n} \{X_i > \nu\}) \leq \sum_{i=1}^{n} P(X_i > \nu) = \sum_{i=1}^{n} (1-F_i(\nu))$ where $F_i(x) = P(X_i \leq x)$. The resulting bound is easily seen to be $P(M_n > \nu) \leq \min(1, \sum_{i=1}^{n} (1-F_i(\nu))$. The important result that Tchen has shown is that there exists a dependent sequence of random variables having the distributions F_i and actually achieving the upper bound. This bound can be applied to the T_i's, using the asymptotic normality of the T_i's. Since the T_i's are strongly dependent, this should provide a conservative bound. These bounds have been included in Figures 3 and 4.

It is interesting to observe that for the Double-Exponential Distribution, the m-Interval Partition Statistic is equivalent to the sign test when the locally most powerful scores are calculated. As mentioned in Section II, the Bivariate m-Interval Partition Statistic is <u>not</u> equivalent to the Bivariate Sign Test since $\max(T_1, \ldots, T_k) \leq \sup T(\theta)$,

$\theta \epsilon [0, 2\pi]$. In fact, if the Type I Error for the Bivariate m-Interval Partition Test via simulation is compared to the null distribution of the Bivariate Sign Test, the latter exceeds the former. Thus the small sample distributions given by Hodges[6] and Klotz[7] provide an upper bound for the performance of the Bivariate m-Interval Partition Statistic for the Bivariate Double-Exponential Distribution.

D. An Application of the Bivariate m-Interval Partition Statistic to Edge Detection

As an example of one of the many applications of this statistic, the problem of edge detection in pattern recognition is considered. Assume that a two-dimensional block pattern with two gray levels, i.e. black and white, is given. The Bivariate m-Interval Partition Statistic is applied to detect and trace the edge of the block pattern. The resulting edge can then possibly be used for feature extraction, classification, or initiation of more sophisticated classification procedures. Other approaches to the edge detection problem may be found in Rosenfeld.[10]

The test pattern illustrated in this paper is a block pattern, i.e. a connected solid pattern centered in a 54x60 array. The points in the solid block pattern have mean 1 and those in the backround have mean zero. It is assumed that at each point of this 54x60 array, 10 random samples are available. That is, there are ten corrupted copies of the test pattern stored. To apply the Bivariate m-Interval Partition Test one forms the bivariate random samples by pairing the samples of adjacent points. Then one can use a sequence of bivariate tests of hypothesis to locate and trace the block pattern edge. The null hypothesis in each of these tests is

H_o: $F(x_i, y_i)$, i.e. (x, y) has mean $(0, 0)$.

The alternative hypothesis H_1 is composite in that it contains any of the following alternatives: (x, y) has mean $(1, 1)$, $(1, 0)$, or $(0, 1)$. Under both the alternative and null hypothesis $F(x, y)$ is symmetric about its mean. For each adjacent pair of points, one can either accept or reject the null hypothesis. For pairs of points in the background, one expects to accept the hypothesis and for pairs of points in the block pattern, one expects to reject the hypothesis. The program used to simulate the edge detection procedure is executed in two stages. The first stage locates an edge point of the block pattern and the second stage traces the edge in a clockwise manner once acquisition of an edge point has been achieved.

The first stage of the program is a pattern search consisting of a sequence of alternating vertical and horizontal jumps which systematically searches the array for the pattern. On horizontal jumps, the pair of points used to test for the presence of the block pattern is taken in the horizontal direction. On vertical jumps, the

corresponding pair is taken in the vertical direction. Once the pattern is located, the program back tracks along the path of the last jump, testing at each point for the edge - i.e. the black-white interface.

The second stage of the processing is the edge tracing. In this stage, pairs of points which are adjacent to the alleged edge point, are tested in a prescribed sequence to determine the next closest edge point. The program traces the edge of the block pattern in a clockwise manner. In this stage of the program, bivariate samples with mean (0, 1) or (1, 0) occur more frequently as the alternative hypothesis. For these alternatives, and error is more likely to occur since the shift of mean is smaller in magnitude than when the alternative is (1, 1). Accordingly, the α and β-curves described in the previous section are based upon an alternative with a shift of mean of the same magnitude, i.e. from (0, 0) to (.707, .707). That is, the threshold used in the simulation was established upon a worst case design criterion. In addition, appropriate coding was inserted in the program to determine if the edge tracing routine was "lost." When this occurred, a branch to a re-initialization program was executed.

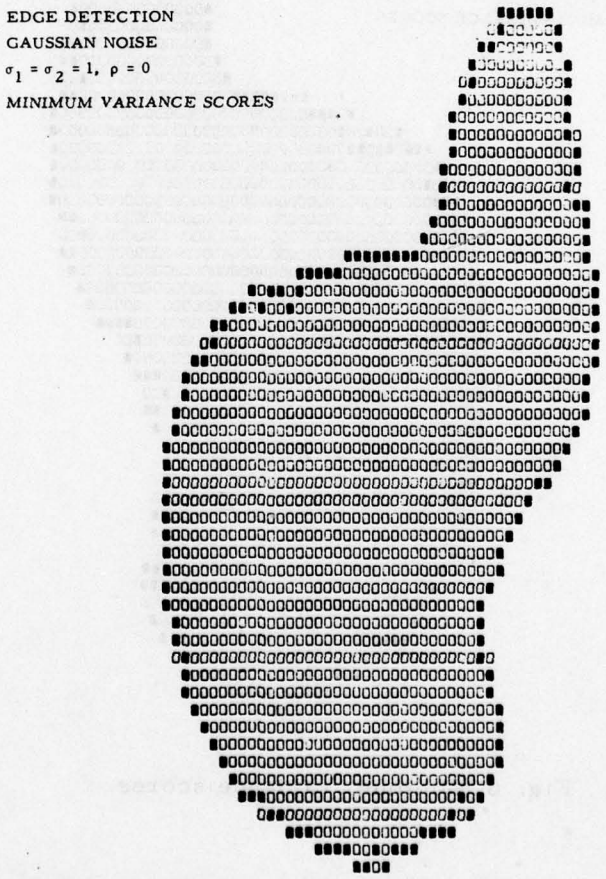


Fig. 5 Minimum variance scores

Simulation runs for the Bivariate Normal, Bivariate Cauchy and Bivariate Double-Exponential were made using both locally most powerful scores and minimum variance scores. Figures 5 and 6 contain illustrations of these runs. In Figure 6, 10% of the noise samples are Bivariate Normal with mean (0, 0) and $\sigma_1^2 = \sigma_2^2 = 10$, $\rho = 0$ and 90% of the samples are Bivariate Normal with mean (0, 0) and $\sigma_1 = \sigma_2 = 1$, $\rho = 0$. The latter distribution provides the noise samples for the simulation illustrated in Fig. 5, with a signal-to-noise ratio of 1. Even at this low signal-to-noise ratio, comparison of Figs. 5 and 6 illustrate the immunity of the Bivariate m-Interval Partition Classifier to Burst Noise.

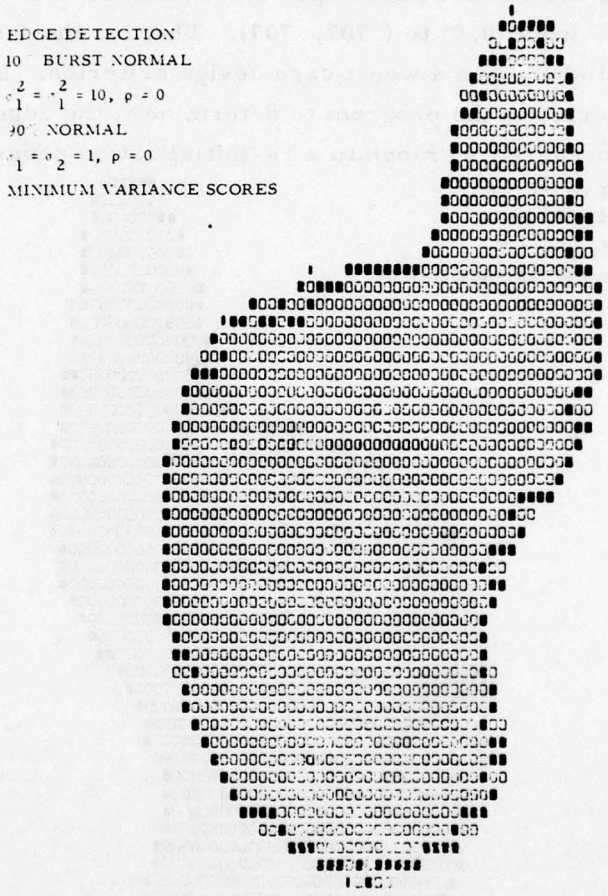


Fig. 6 Minimum variance scores

Extension of this procedure to several gray levels would require adjusting of the threshold of the Bivariate m-Interval Partition Statistic and a different sequence of hypothesis testing if the gray levels are represented by different magnitudes of shift of mean.

Joint Services Technical Advisory Committee
F44620-74-C-0056

P. Kersten and L. Kurz

REFERENCES

1. Y.C. Ching and L. Kurz, "Nonparametric Detectors Based on m-Interval Partitioning," IEEE Trans. on Information Theory, Vol.I T-18, pp. 251-257.
2. J.I. Cochrane and L. Kurz, "Generalized Sign Test Classifiers within a Unified Framework of M-Sample Nonparametric Tests," Proc. Sixth Princeton Conf. Inform. Sc. and Systems, pp. 483-497.
3. M.N. Woinsky and L. Kurz, "An Efficient Nonparametric Detector Based on Three-Sample Classification Model," IEEE Trans. on Info. Theory, Vo. IT-14, pp. 744-752.
4. L. Kurz, "Sequential Nonparametric Two-Way Classification with a Prescribed Maximum Asymptotic Error Probability," Ann. Math. Stat. Vol. 40, pp. 445-455.
5. H. Chadwick and L. Kurz, "Two Sequential Nonparametric Detection Procedures," Information and Control, Vol. 13, pp. 403-428.
6. J.L. Hodges, "A Bivariate Sign Test," Ann. Math. Stat. , Vol. 26, pp. 523-527.
7. J. Klotz, "Null Distribution of the Hodges Bivariate Sign Test," Ann. Math. Stat. Vol. 30, pp. 1029-1033.
8. P. Kersten and L. Kurz, "Bivariate m-Interval Classifiers with Application to Edge Detection," Information and Control (June 1977).
9. A. Tchen, "Inequalities for Distribution with Given Marginals," Technical Report No. 19, Dept. of Stat., Stanford U. (June 1976).
10. A. Rosenfeld, "Picture Processing by Computer,"(New York: Academic Press) pp. 94-102.

A ROBUSTIZED VECTOR RECURSIVE STABILIZER ALGORITHM FOR IMAGE RESTORATION

I. Kadar and L. Kurz

The iterative method of object reconstruction (or image restoration) is reformulated in a new way to find parameters rather than functions. With a block of samples taken from an iterative measurement equation of a time-varying signal in additive noise, the block sample amplitudes become a vector of parameters to be estimated in the presence of noise at each iteration step. This leads directly to the use of the multidimensional extension of Gladyshev's Minimum Variance Least-Squares Stochastic Approximation method (SAMVLS) and, subsequently, to the batch preprocessed Mann-Whitney-Wilcoxon Nonparametric Statistic (MWWNS) robustized vector SAMVLS.

The rebustized SAMVLS algorithm provides immunity to measurement noise outliers in unspecified contaminated noise environments. The method is also both computationally efficient and requires storing of only the last block of batched preprocessed data samples. This represents substantial savings in storage requirements over the direct application of the unrobustized Robbins-Monro Stochastic Approximation (RMSA) method which requires storage of eigenfunctions and has no immunity to noise outliers. The resultant new robust algorithm is equivalent to a natural stabilizer of the measurement-noise-caused instabilities of the iterative method without the need of stopping rule constraints on the recursion.

A. The Parametrized Iterative Object Reconstruction Algorithm

Consider the object reconstruction problem from a truncated measured portion, $g(t)=f(t)p_T(t)$ for which the iterative algorithm described in Ref. 1 becomes (refer to Figure 1)

$$g_{n+1}(t) = [g_n(t) * k(t)]\, p_o(t) + f(t)\, p_T(t),\ n = 1, 2, \ldots \tag{1}$$

where

$$g_1(t) = f(t)\, p_T(t)\ ; \quad p_o(t) \equiv 1 - p_T(t)$$

and $k(t) = (\sin \sigma t)/\pi t$

It is clear that

$$g_{n+1}(t) = \begin{cases} g(t)_1, & |t| < T \\ f_{n+1}(t), & |t| > T \end{cases}$$

In this case the kernel $k(t)$ represents the band-limitedness of $f(t)$, i.e. $f_n(t) = g_{n-1}(t) * k(t)$.

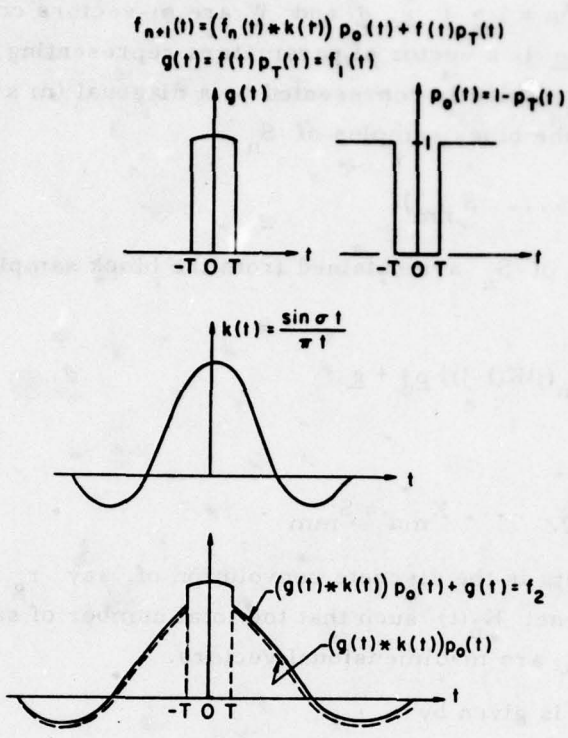


Fig. 1. Plot of quantiles bivariate Cauch-independent components.

The known signal portion $g(t) = f(t)p_T(t)$ in Eq. (1) is replaced with

$$y_n(t) \equiv g(t) + V_n(t) \tag{2}$$

where $V_n(t)$ is an additive noise term, i.i.d for every step n of the iteration. At this point no further assumptions need to be made about the statistics of the noise. At each step of the iteration starting with n=1, a block of samples is taken of all terms in Equation (1). If the functions under considerations are band-limited, then the number of samples needed to represent the functions is defined. Otherwise, one has to assess the significant frequencies of interest and select the sample size to make the aliasing error vanishingly small. Some of these considerations are mentioned in Reference 1 and 2. It will be assumed that even if the sampling rate is greater than the Nyquist rate, the cross-correlation between the block of signal and noise samples is small enough to be negligible and the samples are i.i.d.

Now let $n = 1, 2, 3, \ldots$ and by taking block samples of each time function, one can represent Eq. (1) in terms of "signal," S_n and "noise" $\underline{W}_n$ terms as

$$g_{n+1} = S_n \underline{a} + \underline{W}_n \tag{3}$$

where, say, at some step $n = k \geq 1$, $\underline{g}$, $\underline{\alpha}$ and $\underline{W}$ are m-vectors corresponding to m samples per block where $\underline{\alpha}$ is a vector of parameters representing the amplitude of the time varying signal S_n which is represented by a diagonal (m x m) matrix whose m diagonal elements are the block samples of S_n,

$$S_n = \text{diag}\,(S_{11}, S_{22}, \ldots, S_{mm})$$

where the diagonal entries of S_n are obtained from the block samples of

$$\underline{X}_n \equiv \{[I]_{(mxm)} \sum_{j=0}^{i} g_n(j)K(i-j)\}\, \underline{p}_o + \underline{g} \tag{4}$$

with

$$X_{1n} = S_{11}, \quad X_{2n} = S_{22}, \ldots, X_{mn} = S_{mm} .$$

The sum within the brackets is the discrete convolution of, say, r_g samples of $g_n(t)$ with r_k samples of the kernel $K_n(t)$ such that the total number of samples $m = r_g + r_k - 1$; $\underline{p}_o$ and $\underline{g}$ are m-dimensional vectors.

The noise term $\underline{W}_n$ is given by

$$\underline{W}_n = \{[I]_{(mxm)} \sum_{j=0}^{i} V_n(j)K(i-j)\}\, \underline{p}_o + \underline{V}_n \tag{5}$$

and it is clear from Eqs. (1) and (5) by the discrete convolution operation that any noise perturbation is spread into the reconstructed initially unknown signal portion. It is important to note that the dimension, m, grows with each iteration step n due to the convolution operation which is performed before the recusion for g_{n+1} is iterated. This is not unexpected since the algorith is extrapolating the known signal segment.

It should be noted at this point that the convolution form of the iterative (restoration or extrapolation) algorithm, Eq.(1) is used rather than the FFT implementation suggested by Papoulis[1] and Youla.[3] The convolution form can be adapted directly to the stochastic approximation framework with its similarity to batch preprocessing used in SAMVLS, albeit in this case the convolution is among the block (vector) samples and does not require an initial delay. Furthermore, from a practical point of view charge coupled devices (CCD's) used as analog shift registers are being developed in the 1977-78 time frame for real-time high speed convolution applications for imaging sensor spaceborne signal processing.[4] The convolution operation is performed with CCD's, either by direct storage of the samples and shifting operations,

or by using CCD's to implement the FFT algorithm and form the convolution in a two-step operation. There is no published information available at this time on the relative performance and complexity of the two methods. Both theoretical and experimental work is in progress.[4]

B. The SAMVLS Stabilizer Algorithm

The parametric form of the iterative restoration or extrapolation algorithm $\underline{g}_{n+1} = S_n \underline{\alpha} + \underline{W}_n$ which can be viewed as a measurement equation where, the noise term, $\underline{W}_n$ can be contaminated by outliers directly suggests the application of the vector extension of Gladyshev's theorem to Minimum Variance Least-Squares (SAMVLS) to estimate the time-varying signal parameter, $\underline{\alpha}$ at each step, n.

To apply SAMVLS to the vector parameter estimation problem, one considers each component of the measurement separately and forms, initially, an m-dimensional SAMVLS (since the dimension, m, grows with each iteration step) which can be considered as m scalar SAMVLS algorithms operating in parallel. Specifically, the vector SAMVLS for Eq. (1) becomes

$$\hat{\underline{\alpha}}_{k+1} = \hat{\underline{\alpha}}_k - \frac{1}{k} A_k(\hat{\underline{\alpha}}_1, \ldots, \hat{\underline{\alpha}}_k) [\underline{Y}(\hat{\underline{\alpha}}_k, \underline{\alpha})] \tag{6}$$

where, initially, $\hat{\underline{\alpha}}_k$ is an m-vector, $A_k(\cdot)$ is a diagonal (m x m) adaptive gain matrix. Since $\underline{\alpha}_1$ is arbitrary in the SAMVLS, the optimum choice in this case is to let it equal the amplitude samples of the known signal g(t) for $|t| < T$ and zero for sample values $|t| > T$, making up an (mx1) block sample vector. Actually, the iterative measurement equation (algorithm within the SAMVLS algorithm) is always one step behind as the parameter vector is estimated at step k and the inner algorithm updated. $\underline{Y}(\hat{\underline{\alpha}}_k, \underline{\alpha})$, (mx1), is obtained by correlating the known signal at step k = n with $\underline{g}_{n+1}$,

$$S_k^T \underline{g}_{k+1} = S_k^T(S_k \underline{\alpha} + \underline{W}_k) \quad \text{and} \quad \underline{Y}(\hat{\underline{\alpha}}_k, \alpha) \equiv M(\hat{\underline{\alpha}}_k, \underline{\alpha}) + \underline{Z}(\hat{\underline{\alpha}}_k) .$$

which becomes $\underline{Y}(\hat{\underline{\alpha}}_k, \alpha) = S_k^T \underline{g}_{k+1} - S_k^T S_k \hat{\underline{\alpha}}_k$ where $\hat{\underline{\alpha}}_k$ denotes the current estimate of $\underline{\alpha}$. Substituting Eq. (6) in the expression for $Y(\hat{\underline{\alpha}}_k, \underline{\alpha})$

$$\underline{Y}(\hat{\underline{\alpha}}_k, \underline{\alpha}) = S_k^T [(\underline{\alpha} - \hat{\underline{\alpha}}_k) S_k + \underline{W}_k]$$

and the regression function is given by

$$E\,\underline{Y}(\hat{\underline{\alpha}}_k, \underline{\alpha}) = \underline{M}(\hat{\underline{\alpha}}_k, \underline{\alpha}) = S_k^T S_k(\underline{\alpha} - \hat{\underline{\alpha}}_k)$$

which is linear and has a unique root $\underline{0}$ at $\underline{\alpha} = \hat{\underline{\alpha}}_k$.

As observed previously in the proof of the scalar SAMVLS (see another item by the authors elsewhere in the report), the regression function $\underline{M}(\underline{\hat{\alpha}}_k, \underline{\alpha}) = E\{S_k^T[g_{k+1} - S_k \underline{\hat{\alpha}}_k]\}$ is actually a statement of the multidimensional orthogonality principle, with $[g_{k+1} - S_k \underline{\hat{\alpha}}_k]$ being the "error term" and S_k^T the data. It is well known that the error is orthogonal to the data[5] and $E\{S^T[\underline{g} - S\underline{\hat{\alpha}}]\} = \underline{0}$ and $\underline{\hat{\alpha}}$ minimizes the mean-square error $E\|\underline{g} - S\underline{\hat{\alpha}}\|^2$. This can be interpreted geometrically in terms of orthogonal projections, very much alike to the method of Youla.[3] In reference to Fig. 2, consider the Hilbert space setting of Youla[3] and the orthogonality principle is represented by $(\overline{DE}) \perp (\overline{OE})$ since $(\overline{OD})$ is equivalent to g_{k+1}.

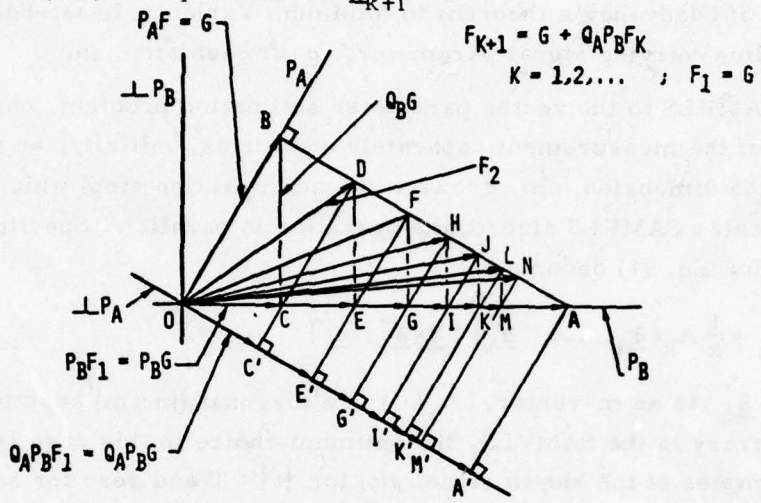


Fig. 2. Minimum variance scores.

Reflecting at this point to the conditions required by Theorem 1 of Ref. 6, the linear regression function satisfies conditions (iv) - (vi) with $M(\underline{\hat{\alpha}}, \underline{\alpha}) = B_k(\underline{\hat{\alpha}} - \underline{\alpha})$ where $B_k = S_k S_k^T$ is a positive definite diagonal $m \times m$ matrix for each k, s.t. $\|B_k\| < \infty$ since one can reasonable assume that the recursion is terminated after a finite number of steps. (With both B_k and A_k diagonal for each, k, $P = I$.) The additive noise term $\underline{Z}(\underline{\hat{\alpha}}) = S_k^T \underline{W}_k$ with $E\,\underline{Z}_k(\underline{\hat{\alpha}}) = \underline{0}$, must have a uniformly bounded variance and have a well-defined covariance matrix as $\underline{\alpha}_k \to \underline{\alpha}$ a. s. This is reflected in condition (vii, a), $\sup_{\underline{\hat{\alpha}}} E\|\underline{Z}(\underline{\hat{\alpha}})\|^{2+2\epsilon} < \infty$ *for some* $\epsilon > 0$ which is satisfied with the Euclidean norm, $\sup_{\underline{\hat{\alpha}}} \{\mathrm{tr}[S_k^T E(\underline{W}\underline{W}^T) S_k]^{1+\epsilon} < \infty$ for each k, with $E(\underline{W}\underline{W}^T)$ non-negative definite. Condition (vii, b) $\lim_{\underline{\hat{\alpha}} \to \underline{\alpha}} E[\underline{Z}(\underline{\hat{\alpha}})\underline{Z}^T(\underline{\hat{\alpha}})] = \pi$, where π is a non-negative definite matrix and where the limit is in the sense of the norm, is satisfied with $\lim_{\underline{\hat{\alpha}} \to \underline{\alpha}} [S_k^T E(\underline{W}\underline{W}^T) S_k] = \pi$. S_k diagonal for each k and $E(\underline{W}\underline{W}^T)$ is a non-negative definite and diagonal by the i. i. d assumption of the problem.

The adaptive gain matrix $A_k(\hat{\underline{\alpha}}_1, \ldots, \hat{\underline{\alpha}}_k)$ needs to be well behaved and a consistent "mean square" estimator of a constant matrix. With $A_k(\cdot)$ diagonal by the assumption of the problem, the eigenvalues of A_k are $a_1^{(k)} > a_2^{(k)} \geq \ldots > a_m^{(k)} > 0$ and the eigenvalues of $B_k = S_k S_k^T$, $B_k = \text{diag}\,(S_{11}^{2(k)}, \ldots, S_{m,m}^{2(k)})$, a diagonal matrix, are $b_1^{(k)} \geq b_2^{(k)} \geq \ldots \geq b_m^{(k)} > 0$. With $a_1 \geq a_2 \geq \ldots \geq a_m > 0$ the eigenvalues of A Condition (viii, a) is satisfied

$$0 < a_1' \leq \inf_{\hat{\underline{\alpha}}} \{\sum_{i=1}^{m} [a_i^{(k)}]^2\}^{\frac{1}{2}} \leq \sup_{\hat{\underline{\alpha}}} \{\sum_{i=1}^{m} [a_i^{(k)}]^2\}^{\frac{1}{2}} \leq a'' < \infty$$

wpl for k large; and $\lim_k a_m^{(k)}(\cdot) \geq a_1^1 > 0$ wpl where A is a constant matrix s.t. $a^1 \leq [\sum_{i=1}^{m} a_1^2]^{\frac{1}{2}} \leq a''$. Condition (ix) $a_1^1 b_m^{(k)} - \epsilon > \frac{1}{2}$ and $a_m b_m^{(k)} - \epsilon > \frac{1}{2}$ for each k, which is required for convergence within the proof.

Theorem 1: Under assumptions (i) - (ix) stated in Theorem of Ref. 6, which were shown to be satisfied above let $r_1^{(k)} > r_2^{(k)} > \ldots > r_m^{(k)} > 0$ be eigenvalues of AB, $B = S_k S_k^T$. Then $k^{\frac{1}{2}}(\hat{\underline{\alpha}}_k - \underline{\alpha})$ is asymptotically normal with mean zero and covariance matrix Q, where Q is a diagonal matrix whose elements are $a_{ii}^2 \pi_{ii}[2a_{ii} S_{ii}^2 - 1]$ where π_{ii} are the elements of $\pi = S^T E(\underline{W}\underline{W}^T) S$ a diagonal matrix, $i = 1, 2, \ldots, m(k)$. For proof see Reference 6.

Comment

It is clear from the above Theorem and from the form of Q, that the asymptotic variance is a function of the power in the signal samples (which are the eigenvalues of B) and the covariance of i.i.d noise vector samples with the dimensionality $m = m(k)$ of Q increasing with each iteration step. The optimum gain coefficient is given by $a_{ii}^{(k)} = 1/(S_{ii}^2)_k$ which minimizes the components of the variance and assures the rate of convergence to be optimal. One has to be careful, however, that $a_{ii}^{(k)} \neq 1/2\,(S_{ii}^2)_k$ since the variance becomes infinite and the recursion diverges. To avoid the instability with $a_{ii}^{(k)} \to 1/2(S_{ii}^2)_k$, one could use the average power in the block signal samples which would rapidly become independent of k and would still guarantee near optimum convergence.

However, even if one could find an estimator for $a_{ii} = 1/S_{ii}^2$ at every step by somehow measuring the energy in the signal samples in the absence of noise, the variance is a function of the covariance of the noise $E(\underline{W}_k \underline{W}_k^T)$, and the variance of the recursion, both asymptotically and in the small sample case, is influenced by the variations in $E(\underline{W}_k \underline{W}_k^T)$. One should recall at this point that the elements of S_k are

$\underline{X}_k = \{[I]_{(m \times m)} \sum_{j=0}^{i} g_k(j) K(i-j)\} \underline{p}_o + \underline{g}$ and the additive noise term is given by

$$\underline{W}_k = \{[I]_{(m \times m)} \sum_{j=0}^{i} V_k(j) K(i-j)\} \underline{p}_o + \underline{V}_k \quad .$$

To alleviate this dependence, on e needs to robustize the SAMVLS algorithm. However, it should be noted that even in the above unrobustized case, the SAMVLS algorithm reduces the mean squared error $E\| \hat{\underline{\alpha}}_k \|^2 = 0(1/k)$ due to the noise in the data, while the iterative measurement equation (algorithm within the SAMVLS algorithm) converges in mean-square as shown by Papoulis.[1] Thus, the reduction of the noise contribution at each iteration step eliminates the instability associated with the iterative measurement algorithm, as long as the rate of convergence of the SAMVLS compensates for the reduction of eigenvalues of the prolate speriodal functions which reduce at a rate depending on the time-bandwidth product, $T\sigma$ with each iteration step.

C. Batch-Nonlinear-Integer Rank Transformation Robustized SAMVLS Image Restoration Algorithm

To robustize, Eq. (6), one introduces batch pre-processing and non-parametric rank statistic of the form (the B-N-L method discussed in another item in this report)

$$W_k^q(\hat{\underline{\alpha}}_k, \underline{\alpha}) = \frac{1}{q^2} \sum_{i=1}^{q} \sum_{j=1}^{q} \operatorname{sgn}\{ S^T_{[i+q(k-1)]} \underline{g}_{[i+qk]} - [S^T S \underline{\alpha}]_{[i+q(k-1)]} \} \tag{7}$$

where $W_k^q(\cdot)$ is an (m(k)x1) vector operator applied component by component. $W_k^q(\cdot)$ is a symmetric version of the Mann-Whitney-Wilcoxon Nonparametric Statistics (MWWNS) with properties summarized here for convenience

$$E[W^q] = 0$$

$$\text{Var } W^q(\cdot) = (2q+1)/3q^2$$

$$\sup_{F,G} W^q = \operatorname{ing}_{F,G} W^q = 1$$

and under H and K

$$\lim_{N \to \infty} P\left[\frac{W^q - EW^q}{\sqrt{\text{Var } W^q}} \leq t \right] = \Phi(t)$$

with asymptotic normality reached with as few as $q = 8$ samples. Furthermore, the above properties do not require symmetry of the CDF, either under H or K.

AD-A050 265 POLYTECHNIC INST OF NEW YORK BROOKLYN MICROWAVE RESE--ETC F/G 20/3
PROGRESS REPORT NO. 42, 15 SEPTEMBER 1976 - 14 SEPTEMBER 1977, --ETC(U)
NOV 77 A A OLINER F44620-74-C-0056
UNCLASSIFIED POLY-MRI-452.42-77 NL

7 OF 7
AD
A050 265



END
DATE
FILMED
3-78
DDC

The robust vector SAMVLS in this case is of the form

$$\underline{\hat{\alpha}}_{k+1} = \underline{\hat{\alpha}}_k - \frac{1}{k} A_k(\underline{\hat{\alpha}}_1, \ldots, \underline{\hat{\alpha}}_k)\, W^q \underline{Y}(\underline{\hat{\alpha}}_k, \underline{\alpha}) \tag{8}$$

which has to satisfy the conditions of Theorem 1. It should be noted here that the batch preprocessing requires an initial delay of q samples during which the algorithm is iterated and the samples are stored. This means that during this period the SAMVLS operates essentially as an unrobustized algorithm and no protection is provided against measurement noise outliers, causing possible instabilities in the "inner" measurement algorithm. Proceeding in a manner similar to the scalar case in Ref. 7, it can be shown that for each m(k) component, say, $\ell = 1, 2, \ldots, m(k)$ with $\underline{M}(\underline{\hat{\alpha}}, \underline{\alpha}) = E\, W^q[\underline{Y}(\underline{\hat{\alpha}}, \underline{\alpha})] = B(\underline{\hat{\alpha}} - \underline{\alpha})$ the diagonal elements of the B-matrix (the slope of the regression function) $B = \text{diag}\,(b_{k_1}, \ldots, b_{k,m(k)}$ in this case are $b_{k\ell} = f_{u_{k\ell} - \nu_{k\ell}}(0)$, $\ell = 1, 2, \ldots, m(k)$ where, $u_{k\ell}$ and $\nu_{k\ell}$ are the batched components of the first and second terms in $\underline{Y}(\underline{\hat{\alpha}}, \underline{\alpha})$, Eq. (7), respectively, and $f_{u_{k\ell} + \nu_{k\ell}}(0)$ is a one-dimensional density corresponding to the ℓth component of $\underline{Y}(\underline{\hat{\alpha}}, \underline{\alpha})$. It is clear from the previous definition of the terms above that $f_{u-\nu}(0)$ is a function of the block signal samples which are time-varying from block to block. This means that the optimum gain coefficient, $A_k = \text{diag}\,(a_{1k}, a_{2k}, \ldots, a_{m(k),k})$ with $a_{i_k} = 1/4\,[f_{u_{ki} - \nu_{ki}}(0)]$ is also time-varying. By assuming that the block signal sample amplitudes can be approximated by an averaged signal level between adjacent blocks, $S_k = S = \text{diag}\,(\text{avg}|S_1|, \ldots, \text{avg}|S_{m(k)}|)$, $f_{u-\nu}(0)$ becomes a constant for each k. Now a simple approximation can be derived by representing $f_{u-\nu}(0)$, $i=1, 2, \ldots, m(k)$ as a generalized Gaussian noise pdf and for a wide class of both thin and heavy tailed pdf's $f_{u_i - \nu_i}(0)$, only varies in the range of 2 to 1. This means that one does not need a very precise estimator of a_{i_k} at each step, k, since the efficiency of the SAMVLS is not very sensitive to changes in a_{i_k}. However, the estimator of a_{i_k} should be robust if one deisres high efficiency independent of the CDF of the measurement noise. Such an estimator is given in Ref. 6, which for each component of $A_k(\cdot)$

$$\hat{a}_{ik} = \frac{q+1}{4(k-1)} \sum_{j=1}^{k-1} \left[Z_{j,[(q/2)+1]} - Z_{j,[q/2]} \right]$$

where $Z_{j,[q/2+1]} = [q/2+1]^{th}$ order statistic form batched component $u_{k\ell}$ of $\underline{Y}(\underline{\hat{\alpha}}, \underline{\alpha})$ and $Z_{j,[q/2]} = [q/2]^{th}$ order statistic from $u_{k\ell}$, $\ell = 1, 2, \ldots, m(k)$, and $[\epsilon]$ defined to be the greatest integer less than or equal to ϵ. It has been shown in Ref. 8 that the

above robust adaptive estimator of the optimum gain coefficient satisfies the conditions of Theorem of Reference 6.

The covariance matrix of the asymptotically normal robustized estimator $k^{\frac{1}{2}}(\hat{\underline{a}}, \underline{a})$ with mean zero is given by $a_{ii}^2 \pi_{ii} [2a_{ii}S_{ii}^2 - 1]$, where $\pi_{ii} = \sigma_{w_{ii}}^2 [I]S^T S$, $i = 1, 2, \ldots, m(k)$, and $\sigma_{w_{ii}}^2 = (2q+1)/3q^2$ which is independent of the measurement noise statistics (compare this result with the covariance of the unrobustized SAMVLS) and the a_{ii}^2 are given by the batched order statistic estimator, defined previously. In this case, the robustness is reached after a small number of iterations and the instability of the recursion is only reached as a limit point as $k \to \infty$.

D. Simulation Results

The SAMVLS method was applied to a signal $f(t) = \sin \sigma t / \pi t$ choosing for T the value $\pi/5\sigma$ in the presence of noise contamination described by the mixture distribution model $f(v) = 0.9n(0,1) + 0.1n(0.8)$. The above example is the same as the one used in Ref. 1 which is used here to illustrate the performance of the "inner" algorithm, Eq. (1), in the absence of noise. This allowed comparing and checking the results with the one in Ref. 1 for the noise-free case using the convolution approach. The result of the simulation of the noise-free "inner" algorithm by the convolution approach is shown in Figure 3. The effect of noise contamination on the convolution implementation of the "inner" algorithm is shown in the same figure indicating divergence. The stable performance of a SAMVLS algorithm in the presence of noise contamination is also shown in Figure 3.

The above simulations were generated using single precision matrix routines. This created a little computational noise in SAMVLS. Therefore, in Fig. 3 a best mean squared fit is shown. If double precision matrix routines are used, the computational noise for all practical purposes disappears.

The robustized version of SAMVLS was not used because of the unavailability of appropriate computer facility and supporting software (advanced IBM370 systems). The simulations were performed on the IBM360/67 Time Sharing computer with CPU-time constraints and slow computational speed.

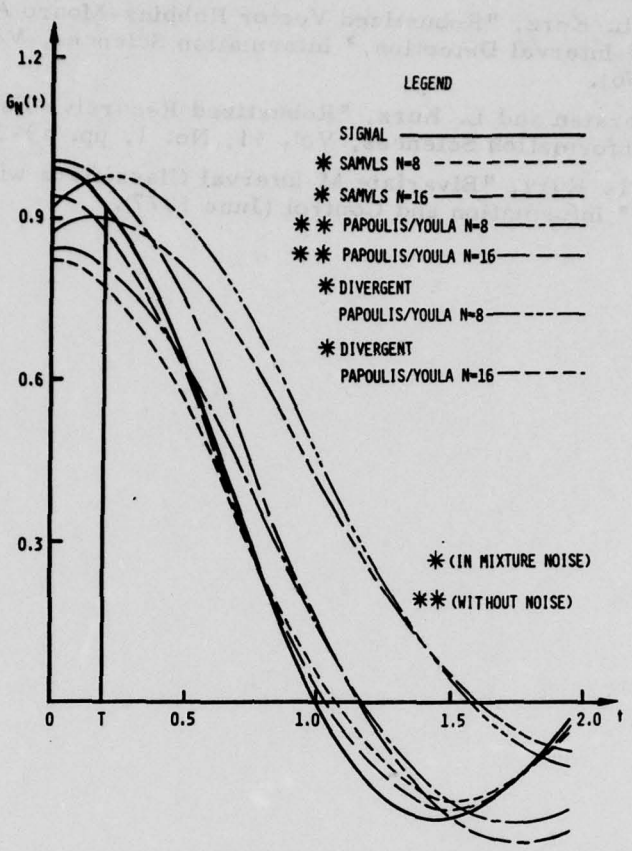


Fig. 3. Plot of Type I error and power vs threshold for the locally most powerful scores N = 500 iterations.

Joint Services Technical Advisory Committee
F44620-74-C-0056

I. Kadar and L. Kurz

Grumman Aerospace Corporation

REFERENCES

1. A. Papoulis, "A New Algorithm in Spectral Analysis and Bandlimited Extrapolation," IEEE Trans. on Circuits and Systems, Vol. CA5-22, No. 9 (September 1975).

2. A. Papoulis, "Systems and Transforms with Applications in Optics," (McGraw-Hill, 1968).

3. D. C. Youla, "Generalized Image Restoration by the Method of Alternating Orthogonal Projections," Progress Report No. 41 to JSTAC, Polytechnic Institute of New York, Report No. R-452.41-76, pp. 544-559 (1976).

4. W. Howle, Private Communication, November 1976 and February, 1977.

5. A. Papoulis, "Probability, Random Variables and Stochastic Processes," (McGraw-Hill, 1965).

6. P. Kersten and L. Kurz, "Robustized Vector Robbins-Monro Algorithm with Application to M-Interval Detection," Information Sciences, Vol. 11, No. 2, pp. 121-140 (1976).

7. J. Evans, P. Kersten and L. Kurz, "Robustized Recursive Estimation with Applications," Information Sciences, Vol. 11, No. 1, pp. 69-92 (1976).

8. P. Kersten and L. Kurz, "Bivariate M-Interval Classifiers with Application to Edge Detection," Information and Control (June 1977).

ESTIMATION OF PROBABILITY DENSITY FUNCTIONS VIA STOCHASTIC APPROXIMATION

P. Kersten and L. Kurz

Tsypkin[1] considers the use of Robbins-Monro (R-M) stochastic approximation procedure to estimate an unknown probability density function. To accomplish this, he approximates the density function f(x) by a finite linear combination of orthonormal functions, i.e.,

$$f(x) = \sum_{i=1}^{N} C_i \phi_i(x) \quad ,$$

where

$$\int \phi_i(x)\,\phi_j(x)\,dx = \delta_{ij} = \langle \phi_i, \phi_j \rangle \quad .$$

The performance measure he minimizes is the integral square error (ISE)

$$I = \int [f(x) - \sum_{i=1}^{N} C_i \phi_i(x)]^2\,dx$$

with respect to C_i, $i = 1, 2, \cdots, N$. Kashyap and Blaydon[2] extended this to general nonorthogonal basis functions. In this report, a set of basis functions is introduced to insure ease of implementation and robustness of the algorithm. It was pointed out previously by the authors[3] that the estimation of the quantiles was an essential ingredient for the implementation of the m-interval detector. Ching and Kurz[4] also remark that the locally most powerful scores for this detector require knowledge of the p.d.f. at the quantiles. They provide one procedure for obtaining these values at the same time that the quantiles are being estimated.

A. Choice of the Basis Functions

Though one can choose basis functions with infinite support, this would require evaluation of each of the basis functions at each of the sample points W_n. An alternative which is computationally more attractive is to use a set of nonorthogonal continuous basis functions with finite support. As an example, consider the following functions, which are merely the square root of a translated Beta p.d.f.

$$\phi(w) = \begin{cases} [(2a+1)!/a!\,a!]^{1/2}(w+\frac{1}{2})^{a/2}(w-\frac{1}{2})^{a/2} & w \in (-\frac{1}{2}, \frac{1}{2}),\ a \text{ even} \\ 0 & \text{otherwise} \end{cases}$$

Here one may construct the basis functions from translations of a single function $\phi(\cdot)$

and vary their interaction merely by choosing the size of the translation appropriately. In general, suppose $\phi_i(x)$ overlaps $\phi_j(x)$ for $j \varepsilon (i-k, \cdots, i, \cdots, i+k)$. If one now minimizes the ISE by taking partial derivatives with respect to C_j, one obtains N coupled linear equations

$$E[\phi_i(w)] = \sum_{j=i-k}^{i+k} C_j \langle \phi_j, \phi_i \rangle \quad , \quad i = 1, \cdots, N, \; C_j = 0 \quad \text{if} \quad j < 1 \quad \text{or} \quad j > N$$

In matrix form

$$E[\Phi(x)] = D\underline{C}$$

where D is a band matrix with $2k+1$ nonzero diagonals given by $d_{ij} = \langle \phi_i, \phi_j \rangle$ and $\vec{C}$ the vector of C_j. The resulting vector recursion equation can be written as

$$\underline{X}_{n+1} = \underline{X}_n - a_n [X_n - D^{-1} \Phi(W_n)]$$

and

$$\underline{X}_{n+1} = \underline{X}_n - a_n [D\underline{X}_n - \Phi(W_n)].$$

where W_n are i.i.d. random samples from the distribution. Since by our choice of basis functions that we specify later, D^{-1} exists, the first version of the recursion relationship is chosen so that the random gain matrix reduces to the identity matrix (see Reference 3). In the notation of Theorem 1 of Ref. 3, $A = B = I = P$. Moreover, $Y(X_n; \vec{C}) = \underline{X}_n - D^{-1} \Phi(W_n)$, $\underline{M}(\underline{X}, \underline{C}) = \underline{X} - \underline{C}$ and accordingly $\underline{Z}_n$ is $\underline{C} - D^{-1} \Phi(W_n)$.

Note that for any sample point only $2k+1$ of the N basis functions are nonzero and need to be evaluated. Moreover, $\partial [\underline{M}(\underline{X}, \vec{C})]_j \,|\, \partial X_j = \delta_{ij}$, so that the gain matrix is "matched" for the optimal rate of convergence of each component of the vector R-M procedure.

B. Results of Simulations

Figures 1 and 2 contain the results of two simulations of the estimation procedure suggested in Section A. Here $N = 13$, $a = 6$, $k = 2$, and

$$\phi_i(w) = \begin{cases} (1.2012)^{1/2} \times 10^2 (w - i/3 + 17/6)^3 \\ \qquad \times (i/3 - 11/6 - w)^3 & , \; i/3 - 17/6 < w < i/3 - 11/6 \\ 0 & \text{otherwise} \end{cases}$$

$$i = 1, \cdots, 13.$$

Simulations were run for both the normal density with mean zero and variance 1/2 and

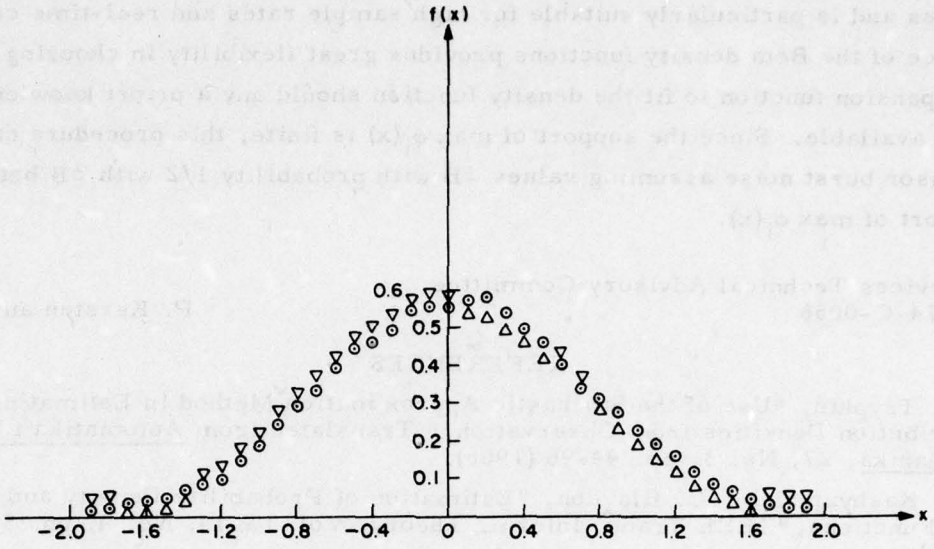


Fig. 1. Probability density function estimation via the vector Robbins-Monro procedure normal distribution function, sample size 1500, $f(x) = \exp(-x^2)/(4\pi)^{1/2}$.

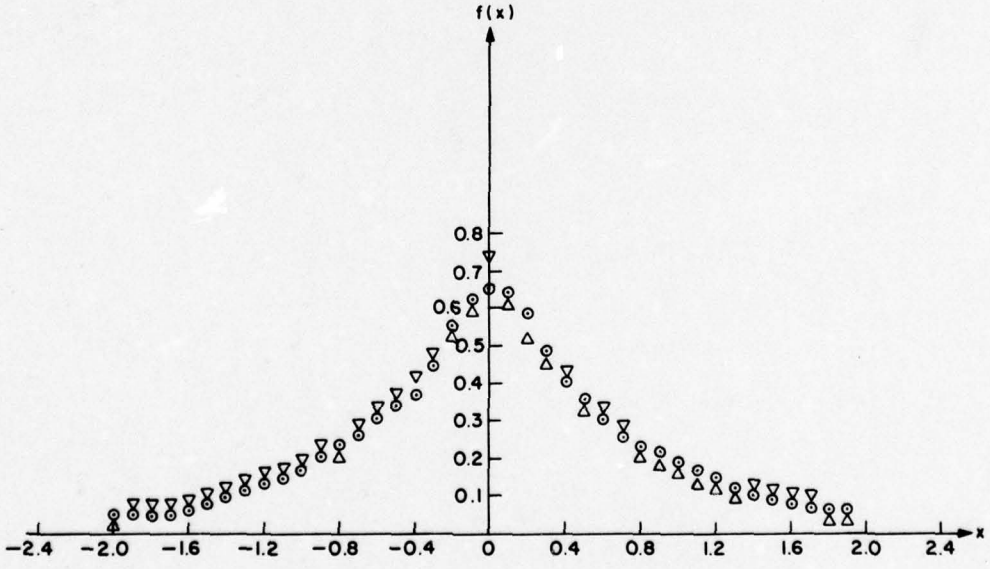


Fig. 2. Probability density function estimation via the vector Robbins-Monro procedure double exponential distribution, sample size 1500, $f(x) = \exp(-|x|\sqrt{2})/\sqrt{2}$.

the Laplace density $f(x) = \exp(-|x|/\lambda)/2\lambda$, $\lambda = 1/\sqrt{2}$. The method requires no storage of samples and is particularly suitable for high sample rates and real-time computation. The choice of the Beta density functions provides great flexibility in choosing the shape of the expansion function to fit the density function should any a priori knowledge of its shape be available. Since the support of $\max_i \phi_i(x)$ is finite, this procedure can be adapted to censor burst noise assuming values $\pm B$ with probability 1/2 with $\pm B$ both outside the support of $\max_i \phi_i(x)$.

Joint Services Technical Advisory Committee
F44620-74-C-0056

P. Kersten and L. Kurz

REFERENCES

1. Y.Z. Tsypkin, "Use of the Stochastic Approximation Method in Estimating Unknown Distribution Densities from Observation," Translated from Automatika i Telemekhanika, 27, No. 3, pp. 94-96 (1966).

2. R.L. Kashyap and C.C. Blaydon, "Estimation of Probability Density and Distribution Functions," IEEE Trans. Inform. Theory, Vol. IT-14, No. 4, pp. 549-556 (1968).

3. P. Kersten and L. Kurz, "Robustized Vector Robbins-Monro Algorithm with Application to M-Interval Detection," Information Sciences, Vol. 11, pp. 121-140 (1976).

4. Y.C. Ching and L. Kurz, "Nonparametric Detectors Based on M-Internal Partitioning," IEEE Trans. Inform. Theory, Vol. IT-18, No. 2, pp. 251-257 (1972).

FIXED POINTS OF RUNNING MEDIANS AND LOCALLY MONOTONIC REGRESSION

S-G. Tyan

A. Introduction

The use of running medians and their combinations as nonlinear smoothers was introduced by Tukey.[1,2] At first they were only recognized as good resistant/robust smoothers against spiky noise or wildly erratic data. Then Rabiner, et al.[3] discovered that these smoothers could also preserve and follow the jump-type discontinuities in the signal sequences, and used them extensively in speech processing. These two properties have made running medians very desirable in areas such as digital image processing,[4,5] exploratory data analysis[1] and speech processing.[3,6] The reason that these median-based smoothers seem to be so promising in engineering application is because they can serve two purposes at the same time: suppressing the spiky noise on the one hand and preserving the jump-type discontinuities in the signal on the other.

In this report running medians are treated as general nonlinear operators and their fixed points are studied. The fixed points are then used as indicators of the effect of running medians on general data sequences. Finally, a few monotonic regression examples are given to justify the use of compound smoothers. All the proofs of the theorems given in this report are omitted, however, they can be found in a paper by the author.[10]

B. Fixed Points of Running Medians

For an infinite sequence x_n, $-\infty < n < \infty$, its running median of odd length $2k+1$ is defined as the sequence y_n, $-\infty < n < \infty$, where y_n is the median of x_i, $n-k \le i \le n+k$. For a finite sequence, several definitions have been used for the end points.[2,7] However, for convenience, we shall use the following:

Definition: Let x_n, $1 \le n \le N$, be a finite sequence. Then its running median of length $2k+1$ is defined as

$$y_n = \text{median of } x_i, \begin{cases} 1 \le i \le 2n-1 & 1 \le n \le k \\ n-k \le i \le n+k, & \text{for } k+1 \le n \le N-k \\ 2n-N \le i \le N & N-k+1 \le n \le N \end{cases}.$$

For a single-side infinite sequence the definition of the running median is clear. From now on we shall use RM(2k+1) to denote running median of length 2k+1 and use RM(2k+1) $\{x_n\}$ for that of a sequence $\{x_n\}$. The following statements, if without further specification, hold for both finite and infinite sequences.

Lemma 1: The running medians are non-expansive mappings, i.e., for any two sequences $\{x_n\}$ and $\{y_n\}$ we have

$$\sup_n |RM(2k+1)\{x_n\} - RM(2k+1)\{y_n\}| \leq \sup_n |x_n - y_n| .$$

Here we have used the supremum of the absolute difference as a measure of distance.

Even though what we actually want to understand is the effects which running medians have on an arbitrary sequence, it seems worthwhile to study those sequences which are unaffected by a running median, namely, its fixed points. This is because of the inherent difficulty with nonlinear operators. On the other hand, fixed points often serve as limits when the operator is used repeatedly and therefore can be used to indicate the trend.

First it is easy to show that a monotonic sequence is invariant under RM(2k+1) for all k. In fact, it can be shown that the requirement on "monotonicity" can be much weakened.

Definition: A sequence $\{x_n\}$ is locally monotonic of length m (abbr. LOMO(m)) iff $\{x_n, x_{n+1}, \ldots, x_{n+m-1}\}$ is monotonic for each n.

Obviously, a LOMO(m) sequence is also LOMO(p) provided $p < m$. Again, assume that $\{x_n\}$ is LOMO(m) and if the segment $\{x_n, x_{n+1}, \ldots, x_{n+m-1}\}$ is increasing, i.e., $x_n < x_{n+m-1}$ and if the segment $\{x_{n+1}, \ldots, x_{n+m}\}$ is decreasing, i.e., $x_{n+1} > x_{n+m}$, then it is apparent that $x_n < x_{n+1} = \ldots = x_{n+m-1} > x_{n+m}$. We have the following lemma:

Lemma 2: If there is any change in trend, then a LOMO(m) sequence must stay constant for at least m-1 samples.

We shall show in Theorem 1 that local monotonicity is enough to guarantee invariance under running medians.

Theorem 1: If a sequence $\{x_n\}$ is LOMO(m) then it is invariant under RM(2k+1) for all k, $k \leq m-2$.

For example, a LOMO(4) sequence is invariant under RM(5). Even though LOMO(k+2) sequences are invariant under RM(2k+1), usually, they do not exhaust the set of fixed points of the RM(2k+1) smoother. It has been found that these fixed points belong to two basically different types. Type I consists of the LOMO(k+2) sequences which have been mentioned.

Theorem 2: If $\{x_n\}$ is a fixed point of RM(2k+1), then $\{x_n\}$ is LOMO(k+2) provided there exists a monotonic segment $\{x_p, x_{p+1}, \ldots, x_{p+k}\}$ of length k+1 .

The above theorem says that if a fixed point of RM(2k+1) is smooth enough (i.e., monotonic) for a segment of length k+1, then it is smooth over the whole length (i.e., LOMO(k+2)).

Definition: A sequence is nowhere LOMO(k), if it does not contain any monotonic segment of length k.

Theorem 3: Let $\{x_n\}$ be a double-side infinite sequence and let it be a fixed point of RM(2k+1). If $\{x_n\}$ is nowhere LOMO(k+1), then $\{x_n\}$ is a duo-valued sequence, i.e., x_n can take on only two values. It is a Type II fixed point.

Example. We will always use 0 and 1 to indicate the two values. For RM(5), the only Type II double-side infinite sequence is ..., 0,1,0,1,... For RM(7), we have also only one double-side infinite Type II sequence which is ...,0,1,1,0,1,0,0,1,0, 1,1,0,1,1,0,0,1,... . Both are periodic. We have some algorithms to generate these Type II sequences for RM(2k+1), however, they do not exhaust the case for k>4. It is not even known whether they are all periodic or not. Since Type II sequences tend to fluctuate faster than Type I sequences of RM(2k+1), they should be considered rather undesirable in a data smoothing problem. If this is true, then they should be suppressed by all means. Be the same reason, the algorithms used to generate them are not presented here.

Theorem 3 does not hold for single-side infinite sequences or finite sequences because of the definition of running medians at the boundaries. However, only a slight modification is necessary.

Theorem 4: Let x_n, $0 \le n \le N$, be a fixed point of RM(2k+1), where $N \ge 2k$. If $\{x_n\}$ is nowhere LOMO(k+1), then it is a duo-valued sequence, say, taking values from $\{0,1\}$, except possibly at the two ends, where $\{x_n\}$ gets away from $\{0,1\}$ monotonically as n approaches toward the two ends 0 and N.

For example, the sequence $\{x_0, x_1, 0, 0, 0, 1, 0, 0, 1, 1, 1, x_{11}\}$ where $x_0 \ge x_1 \ge 1$ and $x_{11} \le 0$, is a fixed point of RM(11). Note that x_n is either 0 or 1 in the central portion of the sequence and that x_n moves away monotonically from $\{0,1\}$ as $n \downarrow 0$ or $n \uparrow N=11$, consequently we always have $x_n(x_n-1) \ge 0$. Since running medians are order statistics, it is obvious that the following lemma is true.

Lemma 3: *If* $\{x_n\}$ *is a fixed point of* RM(2k+1) *and if the mapping* $g(\cdot)$ *is monotonic, then* $\{g(x_n)\}$ *is also a fixed point of* RM(2k+1).

With the above Lemma it is easily seen that, by setting $g(x) = 1$ for $x \ge 1$ and $g(x) = 0$ for $x < 1$, the sequence $\{g(x_n)\}$ is a duo-valued Type II fixed point if $\{x_n\}$ is a Type II finite sequence. On the other hand, we have the following:

Corollary 1: Let $\{x_n\}$, $0 \le n \le N$, be a Type II fixed point of RM(2k+1) taking values from $\{0, 1\}$. Then $\{y_n\}$, $0 \le n \le N$ is also a Type II fixed point of RM(2k+1), where $\{y_n\}$ is constructed according to the following rules:

If $x_i = 1(0)$ for $0 \le i \le n_1$, then $y_i \ge 1$ (≤ 0) and are decreasing (increasing for $0 \le i \le n_1$.

If $x_i = 1(0)$ for $n_2 \le i \le N$, then $y_i \ge i$ (≤ 0) and are increasing (decreasing for $n_2 \le i \le N$.

By Corollary 1, it is clear that all the Type II fixed points of finite length are essentially duo-valued sequences except with slight flexibility at the two ends. They can be constructed from duo-valued sequences immediately by using the method given in the Corollary.

In case $\{x_n\}$ is a single-side infinite sequence, the Theorem 4 and its corollary still hold with two ends replaced with one end.

If the Type II sequences should be bypassed, then we have the converse to Theorem 1.

Theorem 5: (The converse to Theorem 1.) If $\{x_n\}$ is invariant under RM(2p+1) for all $p = 1, 2, \ldots, k$, then $\{x_n\}$ is LOMO(k+2).

Proof: By Theorem 2 and the fact that any sequence is LOMO(2), a fixed point of RM(3), i.e., k=1, must be LOMO(3). Suppose the theorem holds for k-1. Then $\{x_n\}$ which is invariant under RM(2p+1) for all $p = 1, 2, \ldots, k$ must be LOMO(k+1), therefore, each segment of length k+1 is monotonic. By Theorem 2, $\{x_n\}$ is LOMO(k+2). ■

Remark: If a sequence $\{x_n\}$, $0 \le n \le b$, where b may be infinite, is invariant under RM(5), then $x_1 = \text{median}\{x_0, x_1, x_2\}$ which implies that $\{x_0, x_1, x_2\}$ is monotonic. By Theorem 2 the sequence must be LOMO(4). Therefore Type II sequences do not exist for this case.

Rabiner, et al.[3] observed, for a special case, the possible relation between k and m in Theorem 1. It has recently come to our notice that Velleman[8] also described correctly the relation between k and m. He also made the observation that RM(2k+1) tends to create flat tops (or bottoms) of length k+1, which is the very characteristic of LOMO(k+2) sequences. (See Lemma 2.)

Locally monotonic sequences have a certain kind of smoothness in terms of monotonicity: they do not allow any change in trend within a segment of consecutive m samples for LOMO(m) sequences. This excludes the possibility of isolated impulses or a burst of them with duration less than or equal to m-2. On the other hand, jump-type discontinuities are allowed without regard to the magnitude of the jump. Of

course, not all the properties of LOMO sequences are desired. For example, the flats which are necessary for any change in trend can be an eyesore to many viewers.[2,8] On the contrary, the Type II sequences seem to be totally unwanted in a data smoothing problem. Therefore, a good smoother built upon running medians must not have fixed points of this type. In fact, for a smoother T, we should not only consider all the fixed points of T but also those of T^n for all natural numbers n. This is because a sequence, even though not a fixed point of T, may be a recurrent sequence which is just as bad as the Type II sequences. A typical example is the alternating sequence $\ldots, 0, 1, 0, 1, \ldots$ which is also the only Type II sequence of RM(5). The sequence is not a fixed point of RM(3), nevertheless, it is a fixed point of RM(3, 3). Here we use RM(p, q, r) to denote the operation of successively applying RM(p), RM(q) and RM(r) to an arbitrary sequence in this order. We shall also use RM(p * ℓ) to indicate repeating RM(p) for ℓ times.

Since RM(3) and RM(5) are the most frequently used running medians, we shall concentrate on problems related to them exclusively. From Theorem 4 and its Corollary one can see that, even though we have used possibly the simplest definition for the running medians at the boundaries of the finite sequence, the boundary still poses unnecessary complications without yielding much additional information. Therefore in the next section only double-side infinite sequences will be considered.

C. Some Compound Smoothers Related to RM(3)

A class of nonlinear smoothers which also includes RM(3) can be characterized via the following properties. We let $\{z_n\} = T\{x_n\}$ where T is the operator under consideration and let $y_n = \text{median}\{x_{n-1}, x_n, x_{n+1}\}$.

(i) $(z_n - x_n)(z_n - y_n) \leq 0$, (i.e., z_n is between x_n and y_n).

(ii) $y_n \leq z_n < x_n$ or $y_n \geq z_n > x_n$ if $x_n \neq y_n$.

A stronger condition is

(iii) $(z_n - x_n)(z_n - y_n) < 0$ if $x_n \neq y_n$.

Example: RM(3) satisfies (i) and (ii) but not (iii). However, $1/2[I + RM(3)]$, where I is the identity, satisfies (i) and (iii).

Lemma 4: Suppose T satisfies (i). If $\{x_{n-1}, x_n, x_{n+1}\}$ is monotonic, then $z_n = x_n$ and $\{z_{n-1}, z_n, z_{n+1}\}$ is monotonic of the same trend.

Obviously, LOMO(3) sequences are fixed points of any T which satisfies Property (i).

Lemma 5: Let $0 < \alpha < 1$ and let $T_1 * T_2$ be the compound smoother $T_2\{T_1\{x_n\}\}$.

(a) If T_1 and T_2 satisfies (i) then $\alpha T_1 + (1-\alpha)T_2$ and $T_1 * T_2$ both satisfy (i).

(b) If T_1 satisfies (i) and (ii) and if T_2 satisfies (i), then $\alpha T_1 + (1-\alpha)T_2$ satisfies (ii).

(c) If T_1 satisfies (i) and (iii) and if T_2 satisfies (i), then $\alpha T_1 + (1-\alpha)T_2$ satisfies (iii) and $T_1 * T_2$ satisfies (ii). Furthermore, if T_2 also satisfies (iii), then $T_1 * T_2$ satisfies (iii).

Remark: Even if both T_1 and T_2 satisfies (i) and (ii), $T_1 * T_2$ may not satisfy (ii).

Theorem 6. If T satisfies (i) and (ii) and if $\{x_n\}$ is a fixed point of T, then $\{x_n\}$ is a LOMO(3) sequence.

Theorem 7: Let $T = T_1 * T_2$. Suppose T_1 satisfies (i) and (ii) and T_2 satisfies (i). If $\{x_n\}$ is a fixed point of T then $\{x_n\}$ is either a LOMO(3) sequence or an alternating sequence (i.e., $\ldots, 0, 1, 0, 1, \ldots$). If T_1 satisfies (i) and (iii) instead, then $\{x_n\}$ is a fixed point iff it is LOMO(3). The above results hold for $T = T_2 * T_1$.

Remark: Since RM(3) satisfies (i) and (ii), by Theorem 6, it has LOMO(3) sequences as its only fixed points. However, a direct application of Theorem 7 shows that RM(3*ℓ) has both LOMO(3) and alternating sequences as its fixed points if ℓ is even. For ℓ odd, it has only LOMO(3) fixed points. Consequently, the alternating sequence is the only recurrent sequence of RM(3).

Since an alternating sequence is usually considered rather rought, thus it is unwise to use RM(3) or RM(3*ℓ) as the only smoother in smoothing. In the following we consider another kind of arrangement which can forestall this unwanted sequence.

Lemma 6: Let $0 < \alpha < 1$ and let T_1 and T_2 satisfy (i). Then $\{x_n\}$ is a fixed point of $T = \alpha T_1 + (1-\alpha)T_2$ iff it is a fixed point of T_1 and T_2 simultaneously.

Combining Lemma 6 and the remark following Theorem 7 we have

Theorem 8: Let $a_0 \neq 1$, $a_k \geq 0$, and $\sum_{k=0}^{K} a_k = 1$. Then the smoother

$$T = \sum_{k=0}^{K} a_1 \, RM(3*k)$$

has fixed points LOMO(3) sequences only provided for some odd k, $a_k \neq 0$. If $a_k = 0$ for all odd k, then T has both LOMO(3) and alternating sequences as its fixed points.

To obtain the fixed points of T^{ℓ} we have only to rewrite T^{ℓ} as follows:

$$T^{\ell} = T^{\ell-1} * \left[\sum_{k=0}^{K} a_k \, RM(3*k) \right]$$

$$= a_0 T^{\ell-1} + a_1 T^{\ell-1} * RM(3) +$$

$$+ \sum_{k=2}^{K} a_k T^{\ell-1} * RM(3*(k-1)) * RM(3).$$

If $\{x_n\}$ is a fixed point of T^{ℓ}, then by Lemma 6, Theorem 7, and the assumption that $a_0 \neq 1$, it must be a LOMO(3) or an alternating sequence. We can also check that the alternating sequence cannot be invariant under T^{ℓ} unless (1) $a_k = 0$ for all even k and ℓ is even, or (2) $a_k = 0$ for all odd k. In the second case, the alternating sequence is a fixed point of T.

In order to avoid the alternating sequence either as a fixed point or as a recurrent sequence, it seems advisable to use $T_1 = aI + (1-a)RM(3)$ or $T_2 = aRM(3) + (1-a)RM(3*2) = RM(3)*T_1$ in the place of a single RM(3). Indeed, both are free of the alternating sequence. In Section 5 we shall have more discussion on T_1 smoother as well as another smoother with a being a function of $\{x_n\}$ instead of being a constant. The above family of nonlinear smoothers which satisfy property (i) can be viewed as an extension of the RM(3) smoother, nevertheless, they differ from one another in smoothing effect or power. Unfortunately, due to their nonlinear nature, questions concerning details on this aspect do not seem readily amenable to the fixed point theory developed above.

We conclude this section with some miscellaneous results.

Lemma 7: Let $\{y_n\} = RM(3)\{x_n\}$. If (x_{n-1}, x_n, x_{n+1}) is monotonic, then $(y_{n-2}, y_{n-1}, y_n, y_{n+1}, y_{n+2})$ is LOMO(3).

An immediate consequence of Lemma 7 is

Theorem 9: If there exists a monotonic segment $(x_{n-1}\ x_n, x_{n+1})$ then $RM(3*\ell)\{x_n\}$ converges pointwise to a LOMO(3) sequence as $\ell \to \infty$. If $\{x_n\}$ is a finite sequence, $1 \leq n \leq N$, then $RM(3*\ell)\{x_n\}$ is LOMO(3) for some ℓ where $\ell \leq [N-1/2]$.

Another quick result is

Theorem 10: For $T = a + (1-a)RM(3)$, $0 < a < 1$, the sequences $T^{\ell}\{x_n\}$ converge pointwise to a LOMO(3) sequence as $\ell \to \infty$.

So far we have barely mentioned RM(5), one result which we have obtained is

Theorem 11: Let $T = \alpha RM(3) + (1-\alpha)RM(5)$, $0 < \alpha < 1$. Then $\{x_n\}$ is a fixed point iff it is LOMO(4).

D. Locally Monotonic Regression

Assume that the signal sequence $s_1, s_2, \ldots, s_N$ of length N is locally monotonic of length m and that it is corrupted by an error sequence $e_1, e_2, \ldots, e_N$, with e_n i.i.d. random variables. Let $x_n = s_n + e_n$ be the observed sequence. If the error is Gaussian of zero mean, then the maximum likelihood estimate of s_n, $1 \leq n \leq N$, is the sequence which minimizes

$$\sum_{n=1}^{N} (x_n - u_n)^2 \tag{1}$$

among all the LOMO(m) sequences. The special case where m = N and where $\{u_n\}$ is required to be non-decreasing is generally known as isotonic regression or antitonic regression if $\{u_n\}$ is non-increasing. If m = N and if $\{u_n\}$ can be either non-decreasing or non-increasing, then a proper name should be "monotonic regression" as suggested by Barlow, et al. (see Ref. 9, p. 56). Therefore it seems appropriate to call the solution to Eq. (1) as the "locally monotonic regression" of x_n, $1 \leq n \leq N$. At first we were tempted to think that the regression should lie in between the upper and lower envelopes of $\{x_n\}$ (with modifications at the endpoints).[11] However, a simple counterexample shows that it is wrong. In the following, a simpler problem is solved and is used to justify some of the compound smoothers described in Section C. It is straightforward to show that

Lemma 8: Let $(\hat{s}_1, \hat{s}_2, \hat{s}_3)$ be the monotonic regression of (x_1, x_2, x_3). Then $\hat{s}_2 = [x_2 + \text{median}\,(x_1, x_2, x_3)]/2$.

Suppose the signal sequence is LOMO(3) and is coorupted by i.i.d. Gaussian error sequence e_n. Then the maximum likelihood estimate of s_n given x_{n-1}, x_n and x_{n+1} is, according to the above,

$$\hat{s}_n = \frac{1}{2} x_n + \frac{1}{2} \text{ median } \{x_{n-1}, x_n, x_{n+1}\} \tag{2}$$

or

$$\{\hat{s}_n\} = \frac{1}{2}[I + RM(3)]\{x_n\} \tag{3}$$

where I is the identity operator. Equation (3) can be rewritten as

$$\{\hat{s}_n\} = RM(3)\{x_n\} + \frac{1}{2}[I - RM(3)]\{x_n\}$$

which has a structure similar to that of "twicing" as used by Tukey.[2] Here the "rough" $[I - RM(3)]\{x_n\}$ is reduced in magnitude by half and is then added back to the "smooth" $RM(3)\{x_n\}$. (See Beaton and Tukey[7] for the definition of the terms used above.)

If the error has a density function $c \exp(-a|x|^{\gamma})$ then the solution to the simplified regression problem is

$$\hat{s}_n = \begin{cases} (x_n + y_n)/2 & , \quad \gamma > 1 \\ \text{any number between } x_n \text{ and } y_n & , \quad \gamma = 1 \\ x_n \text{ or } y_n & , \quad 0 < \gamma < 1 \end{cases}$$

where $y_n = \text{median}(x_{n-1}, x_n, x_{n+1})$. Apparently, the solution always lies in $I_n = \{x | (x-x_n)(x-y_n) \leq 0\}$.

Another case is the one where the error has Cauchy density, i.e., $1/\pi(1+x^2)$. The solution is

$$\hat{s}_n = \alpha x_n + (1-\alpha) y_n \tag{4}$$

where α is a function of $(x_n - y_n)$ and

$$\alpha = \begin{cases} 1/2 & , \quad |x_n - y_n| \leq 2 \\ 1/2 \pm [1/4 - 1/(x_n - y_n)^2]^{1/2} & , \quad |x_n - y_n| > 2 \end{cases} \tag{5}$$

For smoothing purposes, y_n, which is the median, shall be more reliable than x_n when the difference is large and hence should be given more weight. Therefore the negative sign is chosen in case $|x_n - y_n| > 2$. We may write $\hat{s}_n$ by

$$\hat{s}_n = y_n + (x_n - y_n)\alpha = y_n + g(x_n - y_n)$$

where

$$g(x) = \begin{cases} x/2 & , \quad |x| \leq 2 \\ (1/2 - \sqrt{1/4 - 1/x^2})x & , \quad |x| > 2 \end{cases}$$

The odd function g(x) is linear for small $|x|$ (i.e., $|x| \leq 2$) and it decreases in magnitude to zero as $|x|$ goes beyond that. This is a property which coincides well with our intuition, especially, if the noise has a Gaussian-like center and stretched tails. Again, the nonlinear function g(x) plays the role of Tukey's "smoothing the rough."[2] The design of g(x) can be rather arbitrary to meet different requirements.

The principle is that it should be linear around the origin and decrease in magnitude to zero as $|x|$ goes beyond a certain point.

Since the smoother T of Eq. (4) satisfies Property (iii) of Section C, thus it has only LOMO(3) fixed points. However, when applied to a data sequence, it usually does not generate the flats. A final observation is that $T^{\ell}\{x_n\}$ converges pointwise to a LOMO(3) sequence as $\ell \to \infty$.

Joint Services Technical Advisory Committee
F44620-74-C-0056

S-G. Tyan

REFERENCES

1. J. W. Tukey, "Lecture Notes on Special Topics in Statistics," Statistics 411, (Princeton University, 1974).
2. J. W. Tukey, "Exploratory Data Analysis," (Addison-Wesley, 1977).
3. L. R. Rabiner, et al., "Applications of A Nonlinear Smoothing Algorithm to Speech Processing," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-23, pp. 552-557 (1975).
4. G. W. Wecksung and K. Campbell, "Digital Image Processing at EG&G," *IEEE Computer*, Vol. 7, No. 5, pp. 63-71 (1974).
5. B. R. Frieden," A New Restoring Algorithm for the Preferential Enhancement of Edge Gradients," J. Opt. Soc. Am., Vol. 66, No. 3, pp. 280-283 (1976).
6. N. W. Jayant, "Average- and Median-Based Smoothing Techniques for Improving Digital Speech Quality in the *Presence of Transmission* Errors," IEEE Trans. Comm., Vol. COM-24, pp. 1043-1045 (1976).
7. A. E. Beaton and J. W. Tukey, "The Fitting of Power Series, Meaning Polynomials Illustrated on Band-Spectroscopic Data," Technometrics, Vol. 16, pp. 147-185 (1974).
8. P. F. Velleman, "Robust Non-Linear Data Smoothing," Technical Report No. 89, Series 2, Department of Statistics, (New Jersey: Princeton University, 1975).
9. R. E. Barlow, et al., "Statistical Inference Under Order Restrictions," (New York: John Wiley and Sons, 1972).
10. S. Tyan, "Fixed Points of Running Medians," submitted for publication.
11. S. Tyan, "Data Smoothing *via Order* Statistics," Program Report No. 42 to JSTAC, Polytechnic Institute of New York, Report No. R-452. 42-77 (1977).

MEETING PAPERS

International Conference on Application of Holography and Optical Data Processing, Jerusalem, Israel, August 23-26, 1976

T. Tamir and S.T. Peng, "Network Methods for Integrated Optics Devices"

International Symposium on Systems and Computers, Petras, Greece, August 1976

A. Papoulis, "Computer in Medicine and Astronomy"

A. Papoulis, "Ultrasonics in Medicine: Recent Advances"

A. Papoulis, "Image Reconstruction"

IFAC Conference on Identification and Modelling, Tbilisi, U.S.S.R., September 1976

R. Gran and F. Kozin, "A Non-linear Approach to the Identification of a Non-stationary Earthquake Model"

1976 Annual USNC/URSI Meeting, Amherst, Massachusetts, October 11-15, 1976

H.L. Bertoni and I. Kurland, "Birefringent Prism Couplers"

S. Choudhary and L.B. Felsen, "Modal Propagation in Graded Index Dielectric Waveguides"

S.H. Gross and H. Eun, "Coupling of Atmospheric Waves to Minor Constituent Acoustic Waves"

Second National Conference on Software Engineering, Berkeley, California, October 1976

M.L. Shooman, "Software Error Models - Summary of the State of the Art"

M.L. Shooman, "Structural Models for Software Reliability Prediction"

1976 Ultrasonics Symposium, Annapolis, Maryland, October 1976

F. Freyre and W.C. Wang, "Bragg Interaction of Piezoelectric Leaky Surface Waves and Guided Light in $LiNbO_3$"

Y.L. Hou and H.L. Bertoni, "Variational Approach to Rayleigh Wave Reflection and Transmission at a Liquid Filled Crack"

W.C. Wang, K.C. Whang and L.S. Rosenheck, "Optical Imaging and Memory Using a Si Diode Storage Correlator"

European Conference on Laser Plasma Fusion, Palaiseau, France, October 1976

E.A. Mishkin, "Radiation Losses in a Thermonuclear Reaction"

IEEE Conference on Communications and Power, Montreal, Canada, October 21-22, 1976

T.S. Sundresh, F.A. Cassara and H. Schachter, "Suppression of Interchannel Interference in FM Receivers"

1976 ORSA/TIMS Joint National Meeting, Miami Beach, Florida, November 1976

M.L. Shooman and H. Ruston, "Cost Reducing, High Reliability Programming Techniques"

Pulsed Power Conference, Lubbock, Texas, November 1976

E.E. Kunhardt, J. Tardiff and B.R. Cheo, "The Bouncing Conductor Generator"

1976 IEEE National Telecommunications Conference, Dallas, Texas, November 1976

A. Livne and R.R. Boorstyn, "On a Technique for Adaptive Routing"

R.J. Pennotti and R.R. Boorstyn, "Channel Assignments for Cellular Mobile Telecommunications Systems"

18th Annual Meeting of the Division of Plasma Physics, American Physical Society, San Francisco, California, November 1976

E.S. Cassedy, "Time-Dynamic Growth of Multimode Parametric Oscillations"

J. Faehl, E.S. Cassedy and W.L. Kruer, "Effect of Self-Consistent Density Fluctuation on a Standing Wave Pumped Parametric Instability"

1977 Annual Reliability and Maintainability Symposium, Philadelphia, Pennsylvania, January 1977

M.L. Shooman and S. Sinkar, "Generation of Reliability and Safety Data by Analysis of Expert Opinion"

Electromagnetics Colloquium, University of Mississippi, February 1977

N. Marcuvitz, "Nonlinear and Turbulent Wave Propagation"

Plasma Physics Seminar, Naval Research Laboratory, Washington, DC February 1977

N. Marcuvitz, "*Theory of Plasma* Turbulence"

Bell Telephone Laboratories Seminar, Holmdel, New Jersey, February 1977

R.R. Boorstyn, "The Design of Centralized Network Topologies, Constrained Spanning Trees and Matroids"

L.B. Felsen, "Rays, Modes and Beams in Optical Fiber Waveguides"

1977 Meeting of the American Physical Society, San Diego, California, March 1977

D. Lischner and H.J. Juretschke, "Origin of the Metallic Field Effect"

P. Mazumdar and H.J. Juretschke, "Oscillatory Metallic Field Effect in Thin Ferromagnetic Films"

Conference on Transportation, University of Michigan, April 1977

C. Berger and L. Shaw, "Discrete Event Simulation of Freeway Traffic"

Collective Phenomena Seminar, Moscow, U.S.S.R., April 18-20, 1977

N. Marcuvitz, "Quasiparticle Descriptions in Collective Systems"

Electrochemical Society 151st Meeting, Philadelphia, Pennsylvania, May 8-13, 1977

W.C. Wang, H. Schachter, F.A. Cassara, K.C. Whang and L. S. Rosenheck, "Information Storage Using Diode Arrays SAW Correlator"

1977 IEEE Acoustics, Speech and Signal Processing Conference, Hartford, Connecticut, May 9-11, 1977

H. Schachter, F.A. Cassara and L.S. Rosenheck, "Stability and Noise Analysis of the Acoustic-Electric Phase-Locked Loop Device"

IEEE International Conference on Plasma Science, Troy, New York, May 23-25, 1977

H. Kudyan and K. Chung, "Evolution of Transverse Instability in a Hollow Cylindrical Weakly-Ionized Plasma Column"

S.P. Kuo and B.R. Cheo, "Parametric Decay Instabilities in a Uniform Magneto Plasma"

S.P. Kuo and B.R. Cheo, "Saturation of Parametric Instability due to Anamalous Diffusion Process"

T.Q. Yip, S.P. Kuo and B.R. Cheo, "Experimental Study of Time Evolution of Parametrically Excited Instabilities in a Magneto Plasma"

1977 Meeting of the American Geophysical Union, Washington, DC, May 30-June 3, 1977

H. Eun and S.H. Gross, "The Response of Minor Species to Gravity Waves in the Thermosphere"

G.V. Ramanathan and S.H. Gross, "Theoretical Investigation of Dynamic Outflow from Planetary Atmospheres"

Fourth Rochester Conference on Coherence and Quantum Optics, Rochester, New York, June 8-10, 1977

F.P. Mattar and M.C. Newstein, "Can Coherent Effects be Important After All?"

D. Scarl and S. McAfee, "Two-photon Time Correlations in Diverging Spatially Coherent Light"

International Magnetics Conference, Los Angeles, California, June 6-9, 1977

E. Levi, L. Birenbaum and Z. Zabar, "Concerning the Design of Inductor Synchronous Motors Fed by Current Source Inverters"

1977 Joint Automatic Control Conference (IFAC), San Francisco, California, June 1977

F. Kozin and S. Sugmoto, "Determining Stability Properties of Stochastic Systems from Sample Observations"

M.L. Shooman, "Need and Use of Analytical Methods in Reliability"

M.L. Shooman, "The Spectre of Software Reliability and Its Exorcism"

1977 International AP-S Symposium and URSI International Electromagnetic Symposium, Stanford, California, June 1977

R.S. Chu, J.A. Konj and T. Tamir, "The Diffraction of Gaussian Beams by Periodic Layers"

S.T. Peng and A.A. Oliner, "Waveguides for Integrated Optics: Comparison of Analysis Methods and New Physical Effects"

S.T. Peng, T. Tamir and K.C. Cheng, "Leaky Wave Characteristics of Dielectric Gratings with Arbitrary Profiles"

World Electrotechnical Congress, Moscow, U.S.S.R., June 21-26, 1977

E. Levi, "Linear Electric Propulsion"

International Conference on Integrated Optics and Optical Fiber Communication, Tokyo, Japan, July 1977

S.T. Peng and A.A. Oliner, "Leakage and Resonance Effects on Strip Waveguides for Integrated Optics"

Symposium on Computer Methods in Engineering, University of Southern California, August 1977

F. Kozin, "Estimation and Model for Non-stationary Time Series"

Seventh European Microwave Conference, Copenhagen, Denmark, September 1977

H.L. Bertoni, A.C. Green and L.B. Felsen, "Shadowing of an Inhomogeneous Plane Wave by an Edge"

A.A. Oliner and S.T. Peng, "Leakage and Resonance Effects on Guiding Structures for Optical and Millimeter Waves"

Sixth World Conference on Earthquake Engineering, New Delhi, India, 1977

R.F. Drenick and P.C. Wang, "Critical Seismic Excitation and Response of Structures"

H. Kano and R.F. Drenick, "The Critical Excitation of Inelastic Structures"

JOURNAL ARTICLES AND CORRESPONDENCE

S.R. Barone, M.A. Narcowich and F.J. Narcowich, "Floquet Theory and Applications," Phys. Rev. A, 15, No. 3, 1109-1125 (March 1977).

H.L. Bertoni, A.C. Green and L.B. Felsen, "Shadowing of an Inhomogeneous Plane Wave by an Edge," to be published in J. Opt. Soc. Am.

J.J. Bongiorno, Jr. and D.C. Youla, "On the Design of Single-Loop Single Input-Output Feedback Control Systems in the Complex-Frequency Domain," IEEE Trans. Automatic Control, AC-22, No. 3 (June 1977).

F.A. Cassara, "FM Noise Clicks of the Second Kind in the Frequency Demodulator with Feedback," Alta Frequenza, XLV, No. 10, 595-599 (October 1976).

F.A. Cassara, "Experimental Optimization of the Signal-to-Noise Ratio Response of the Frequency Demodulator with Feedback in the Presence of Interchannel Interference," Int. J. Electr., 42, No. 4, 369-375 (April 1977).

F.A. Cassara, "A Laboratory Experiment on Communication Electronics," to be published in Int. J. Electr. Eng. Ed.

F.A. Cassara and B. Senitzky, "A Freshman General Engineering Lab," IEEE Trans. on Education, 19, No. 4, 139-142 (November 1976).

E.S. Cassedy, "Purely-growing Parametric Instabilities in Plasmas at Near Critical Densities," J. Plasma Phys. (Gr. Brit.), 16, Part 3, 371-388 (1976).

B.R. Cheo, "Truncation of Infinite Sheath Spiral Antennas," IEEE Trans. Antennas Propag., AP-25, No. 3, 359-365 (May 1977).

R.S. Chu and T. Tamir, "Diffraction of Gaussian Beams by Periodically Modulated Media for Incidence Close to a Bragg Angle," J. Opt. Soc. Am., 66, 1438-1440 (December 1976).

K. Chung and H. Eun, "Observation of Microwave Mixing in Air Glow Plasma," IEEE Trans. on Plasma Science, PS-5, 41-48 (1977).

M. Eschwei and S. Gottfried, "A Sputter-Etch Plus Evaporated Film Technique for Inexpensive Reliable Ohmic Contacts to Silicon Devices," to be published in J. Vac. Sci. Technol.

H. Eun and S.H. Gross, "Ionispheric Disturbances and Gravity Waves," J. Geophys. Res., 81, 3261-3270 (1976).

H. Eun and S.H. Gross, "Resonant Coupling of Ionization Waves and Acoustic Gravity Waves in the Presence of a Magnetic Field," J. Geophys. Res., 81, 5964-5970 (1976).

H. Frank and R.R. Boorstyn, "Large-Scale Network Topological Optimization," IEEE Trans. Comm., COM-25, No. 1, 29-48 (January 1977).

Y. Fujimoto and E.A. Mishkin, "Hydrodynamics of an Imploding Small Sphere," Bull. Am. Phys. Soc., 1103 (October 1976).

H.M. Gibbs, B. Bolger, F.P. Mattar, M.C. Newstein, G. Forster and P.E. Toscheck, "Coherent On-Resonance Self-Focusing of Optical Pulses in Absorbers," Phys. Rev. Lett., 37, 1743-1746 (December 1976).

M. Greenblatt and E. Banks, "CdF_2: YbF_3 ErE_3 – An Efficient Infrared to Visible Upconverting System," J. Electrochem. Soc., 124, 409 (1977).

M. Greenblatt, J.H. Pifer and E. Banks, "Electron Spin Resonance of CiO_4^{3-} in Fluoropatite $Ca_5(PO_4)$ $F.^4$," J. Chem. Phys., 66, 559 (1977).

N.N. Greenwood, F. Viegas, E. Banks and M. Nemiroff, "The Defect Model and Oxidation State of Europium and *Molybdenum in* Eu_x MoO_4," Inorg. Chem., 15, 2317 (1976).

S.H. Gross and H. Eun, "Traveling Neutral Disturbances," Geophys. Res. Lett., 3, 257-260 (1976).

S.H. Gross and G.V. Ramanathan, "The Atmosphere of Io," Icarus, 29, 493-502 (1976).

H.J. Juretschke, "Simple Derivation of the Maxwell Stress Tensor and Electrostrictive Effects in Crystals," Am. J. Phys., 45, 277 (1977).

P. Kersten and L. Kurz, "Robustized Vector Robbins-Monro Algorithm with Applications to M-Interval Detection," Info. Sci., 11, 121-140 (1976).

P. Kersten and L. Kurz, "Bivariate m-Interval Classifiers with Application to Edge Detection," Information and Control, 34, No. 2, 152-168 (June 1977).

F. Kozin and T.S. Lee, "Almost Sure Asymptotic Likelihood Theory for Diffusion Processes," to be published in J. Appl. Probability.

H. Kudyan, "On the Interpretation of Electrostatic Energy Analyzer Data of a Flowing Plasma," to be published in Rev. Sci. Instrum.

E.E. Kunhardt and B.R. Cheo, "Observation of Wave Packet Bifurcation in a Magneto Plasma Column," Phys. Rev. Lett., 37, No. 25, 1688-1691 (December 1976).

E.E. Kunhardt and B.R. Cheo, "Propagation of Non-linear Waves Along a Magneto Plasma Column," to be published in Phys. of Fluids.

I.J. Kurland and H.L. Bertoni, "Birefringent Prism Couplers for Thin-Film Optical Waveguides," to be published in Appl. Opt.

E. Levi, L. Birenbaum and Z. Zabar, "Concerning the Design of Inductor Synchronous Motors Fed by Current Source Inverters," IEEE Trans. on Magnetics, MAG-13, No. 5, 1421-1423 (September 1977).

R.C.M. Li, A.A. Oliner and H.L. Bertoni, "Microwave Network Analysis of Surface Acoustic Waveguides, I: Flat Overlay Guides," IEEE Trans. Sonics Ultrason., SU-24, No. 2, 66-78 (March 1977).

S. Markman, R.C.M. Li, A.A. Oliner and H.L. Bertoni, "Microwave Network Analyses of Surface Acoustic Waveguides, II: Rectangular Ridge Guides," IEEE Trans. Sonics Ultrason., SU-24, No. 2, 79-87 (March 1977).

E.A. Mishkin, "Brightness of a Small Compressed Pellet," Bull. Am. Phys. Soc., 1083 (October 1976).

C. Mohwinkel and L. Kurz, "Computer Picture Processing and Enhancement by Localized Operations," Computer Graphics and Image Processing, 5, 401-424 (1976).

M.C. Newstein and F.P. Mattar, "Transverse Effects Associated with the Propagation of Coherent Optical Pulses in Resonant Media," to be published in IEEE J. Quantum Electron.

A.A. Oliner, "Acoustic Surface Waveguides and Comparisons with Optical Waveguides," IEEE Trans. Microwave Theory Tech., MTT-24, No. 12, 914-920 (December 1976).

G.R. Osche, " Dirac and Dirac-Pauli Equations in the Foldy-Wouthuysen Representation," Phys. Rev. D, 15, No. 8, 2181-2185 (April 1977).

A. Papoulis, "The Two-To-One Rule in Data Smoothing," to be published in IEEE Trans. Inf. Theory.

A. Papoulis, "The Problem of Transmission Zeros in Deconvolution," to be published in IEEE Trans. Inf. Theory.

A. Papoulis, "System Identification with Non-stationary Inputs," to be published in IEEE Trans. Inf. Theory.

D.C. Rife and R.R. Boorstyn, "Multiple Tone Parameter Estimation from Discrete-Time Observations," Bell System Technical Journal, 55, No. 9, 1387-1410 (November 1976).

B. Rudner, book review of "Signal Processing: Discrete Spectral Analysis, Detection and Estimation," by M. Schwartz and L. Shaw, J. of the Franklin Institute, 303, No. 2, 219-220 (February 1977).

D.B. Scarl and S.R. Smith, "Two-photon Correlations in Light From a Gain Tube," Phys. Rev. A, 14, 2355-2356 (December 1976).

L. Shaw, "Busy Period Control of Queues Based on Waiting Times at Arrivals," Eviata, 25, No. 3, 538 (May 1977).

L. Shaw and M.L. Shooman, "Confidence Bounds and Propagation of Uncertainties in System Availability and Reliability Computations," Naval Research Logistics Quarterly, 23, No. 4, 673-685 (December 1976).

T. Tamir, "Radio Wave Propagation Along Mixed Paths in Forest Environments," to be published in IEEE Trans. Antennas Propag.

Y. Thomas, D. Sarlat and L. Shaw, "Receding Horizon Approach to the Synthesis of Nonlinear Multivariate Regulators," Electron. Lett., 13, No. 11, 329-330 (May 1977).

W.C. Wang, "On the Physical Process of SAW Schottky Diode Memory Correlator," Electron. Lett., 12, 105 (1976).

N. Wattanapanom and L. Shaw, "Optimal Inspection Schedules for Failure Detection in a Model where Tests Hasten Failures," to be published in Operations Research.

M.M. Zaret, W.Z. Snyder and L. Birenbaum, "Cataract After Exposure to Non-Ionizing Radiant Energy," Brit. J. Ophthalmology, 60, No. 9, 632-637 (September 1976).

PATENTS

T. Tamir and S.T. Peng, Directional Radiation by Asymmetrical Dielectric Gratings, U.S. Patent No. 3982810, September 1976.

RECENT BOOKS

H. Ruston, "Programming with PL/I," (New York: McGraw Hill, 1977).

CONTRIBUTIONS TO RECENT BOOKS AND TO PROCEEDINGS

C. Berger and L. Shaw, "Discrete Event Simulation of Freeway Traffic," Digest of Papers presented at Conference on Transportation, University of Michigan (April 1977).

H. L. Bertoni and I. Kurland, "Birefringent Prism Couplers," Proc. of USNC/URSI Annual Meeting, Amherst, Massachusetts (October 1976).

H. L. Bertoni, A. C. Green and L. B. Felsen, "Shadowing of an Inhomogeneous Plane Wave by an Edge," Proc. of Seventh Microwave Conference, Copenhagen, Denmark (September 1977).

R. R. Boorstyn, "The Design of Centralized Network Topologies Constrained Spanning Trees and Matroids," Digest of Papers presented at Bell Telephone Laboratories Seminar, Holmdel, New Jersey (February 1977).

E. S. Cassedy, "Time-Dynamic Growth of Multimode Parametric Oscillations," Proc. of 18th Annual Meeting of the Division of Plasma Physics, American Physical Society, San Francisco, California, paper 10F1 (November 1976).

S. Choudhary and L. B. Felsen, "Modal Propagation in Graded Index Dielectric Waveguides," Proc. of USNC/URSI Annual Meeting, Amherst, Massachusetts (October 1976).

R. S. Chu, J. A. Konj and T. Tamir, "The Diffraction of Gaussian Beams by Periodic Layers," Proc. of 1977 International AP-S Symposium and URSI International Electromagnetic Symposium, Stanford, California (June 1977).

R. F. Drenick, "Organization and Control," chapter in New Directions, Y. C. Ho, Ed. (New York: Plenum Press, 1976).

R. F. Drenick and P. C. Wang, "Critical Seismic Excitation and Response of Structures," Proc. of Sixth World Conference on Earthquake Engineering, New Delhi, India (1977).

H. Eun and S. H. Gross, "The Response of Minor Species to Gravity Waves in the Thermosphere," Proc. of 1977 Meeting of the American Geophysical Union, Washington, DC (May 1977).

J. Faehl, E. S. Cassedy and W. L. Kruer, "Effect of Self-Consistent Density Fluctuation on a Standing Wave Pumped Parametric Instability," Proc. of 18th Annual Meeting of the Division of Plasma Physics, American Physical Society, San Francisco, California, paper 8H2 (November 1976).

L. B. Felsen, "Rays, Modes and Beams in Optical Fiber Waveguides," Digest of Papers presented at Bell Telephone Laboratories Seminar, Holmdel, New Jersey (February 1977).

F. Freyre and W. C. Wang, "Bragg Interaction of Piezoelectric Leaky Surface Waves and Guided Light in $LiNbO_3$," Proc. of 1976 IEEE Ultrasonics Symposium, Annapolis, Maryland, 193-196 (October 1976).

R. Gran and F. Kozin, "A Non-linear Approach to the Identification of a Non-stationary Earthquake Model," Proc. of IFAC Conference on Identification and Modelling, Tbilisi, U.S.S.R. (September 1976).

S.H. Gross and H. Eun, "Coupling of Atmospheric Waves to Minor Constituent Acoustic Waves," Proc. of USNC/URSI Annual Meeting, Amherst, Massachusetts (October 1976).

Y.L. Hou and H.L. Bertoni, "Variational Approach to Rayleigh Wave Reflection and Transmission at a Liquid Filled Crack," Proc. of 1976 IEEE Ultrasonics Symposium, Annapolis, Maryland (October 1976).

H. Kano and R.F. Drenick, "The Critical Excitation of Inelastic Structures," Proc. of Sixth World Conference on Earthquake Engineering, New Delhi, India (1977).

F. Kozin, "Estimation and Model for Nonstationary Time Series," Proc. of Symposium on Computer Methods in Engineering, University of Southern California (August 1977).

F. Kozin and S. Sugmoto, "Determining Stability Properties of Stochastic Systems from Sample Observations," Proc. of Joint Automatic Control Conference (IFAC), San Francisco, California (June 1977).

H. Kudyan and K. Chung, "Evolution of Transverse Instability in a Hollow Cylindrical Weakly-Ionized Plasma Column," Proc. of IEEE International Conference on Plasma Science, Troy, New York (May 1977).

S.P. Kuo and B.R. Cheo, "Parametric Decay Instabilities in a Uniform Magneto Plasma," Proc. of IEEE International Conference on Plasma Science, Troy, New York (May 1977).

S.P. Kuo and B.R. Cheo, "Saturation of Parametric Instability due to Anomalous Diffusion Process," Proc. of IEEE International Conference on Plasma Science, Troy, New York (May 1977).

E. Levi, "Linear Electric Propulsion," Proc. of World Electrotechnical Congress, Moscow, U.S.S.R. (June 1977).

E. Levi, L. Birenbaum and Z. Zabar, "Concerning the Design of Inductor Synchronous Motors Fed by Current Source Inverters," Proc. of International Magnetics Conference, Los Angeles, California (June 1977).

D. Lischner and H.J. Juretschke, "Origin of the Metallic Field Effect," Proc. of Meeting of the American Physical Society, San Diego, California, 22, 453 (Bull. Am. Phys. Soc., 1977).

A. Livne and R.R. Boorstyn, "On a Technique for Adaptive Routing," Proc. of IEEE National Telecommunications Conference, Dallas, Texas (December 1976).

N. Marcuvitz, "Theory of Plasma Turbulence," Proc. of Plasma Physics Seminar, Naval Research Laboratory, Washington, DC (February 1977).

N. Marcuvitz, "Nonlinear and Turbulent Wave Propagation," Proc. of Electromagnetics Colloquium, University of Mississippi (February 1977).

N. Marcuvitz, "Quasiparticle Descriptions in Collective Systems," Proc. of Collective Phenomena Seminar, Moscow, U.S.S.R. (April 1977).

F.P. Mattar and M.C. Newstein, "Can Coherent Effects be Important After All?" Proc. of Fourth Rochester Conference on Coherence and Quantum Optics, Rochester, New York (June 1977).

P. Mazumdar and H.J. Juretschke, "Oscillatory Metallic Field Effect in Thin Ferromagnetic Films," Proc. of Meeting of the American Physical Society, San Diego, California, 22, 453 (Bull. Am. Phys. Soc., 1977).

E.A. Mishkin, "Radiation Losses in a Thermonuclear Reaction," Proc. of European Conference on Laser Plasma Fusion, Palaiseau, France (October 1976).

A.A. Oliner and S.T. Peng, "Leakage and Resonance Effects on Guiding Structures for Optical and Millimeter Waves," Proc. of Seventh European Microwave Conference, Copenhagen, Denmark (September 1977).

A. Papoulis, "Computer in Medicine and Astronomy," Proc. of International Symposium on Systems and Computers, Petras, Greece (August 1976).

A. Papoulis, "Ultrasonics in Medicine: Recent Advances," Proc. of International Symposium on Systems and Computers, Petras, Greece (August 1976).

A. Papoulis, "Image Reconstruction," Proc. of International Symposium on Systems and Computers, Petras, Greece (August 1976).

S.T. Peng and A.A. Oliner, "Waveguides for Integrated Optics: Comparison of Analysis Methods and New Physical Effects," Proc. of 1977 International AP-S Symposium and URSI International Electromagnetic Symposium, Stanford, California (June 1977).

S.T. Peng and A.A. Oliner, "Leakage and Resonance Effects on Strip Waveguides for Integrated Optics," Proc. of International Conference on Integrated Optics and Optical Fiber Communication, Tokyo, Japan (July 1977).

S.T. Peng, T. Tamir and K.C. Cheng, "Leaky-Wave Characteristics of Dielectric Gratings with Arbitrary Profiles," Proc. of 1977 International AP-S Symposium and URSI International Electromagnetic Symposium, Stanford, California (June 1977).

R.J. Pennotti and R.R. Boorstyn, "Channel Assignments for Cellular Mobile Telecommunications Systems," Proc. of IEEE National Telecommunications Conference, Dallas, Texas (December 1976).

G.V. Ramanathan and S.H. Gross, "Theoretical Investigation of Dynamic Outflow from Planetary Atmospheres," Proc. of 1977 Meeting of the American Geophysical Union, Washington, DC (May 1977).

S.W. Rosenthal, chairman of session on "Biological Effects of Microwaves – Medical and Biological Applications," Proc. of 1977 International Microwave Power Institute's Microwave Power Symposium, Minneapolis, Minnesota (May 1977).

D. Scarl and S. McAfee, "Two-photon Time Correlations in Diverging Spatially Coherent Light," Proc. of Fourth Rochester Conference on Coherence and Quantum Optics, Rochester, New York (June 1977).

H. Schachter, F.A. Cassara and L.S. Rosenheck, "Stability and Noise Analysis of the Acousto-Electric Phase-Locked Loop Device," Proc. of 1977 IEEE Acoustics, Speech, and Signal Processing Conference, Hartford, Connecticut, 389-392 (May 1977).

M.L. Shooman, "Software Error Models–Summary of the State of the Art," Proc. of Second National Conference on Software Engineering, Berkeley, California (October 1976).

M.L. Shooman, "Structural Models for Software Reliability Prediction," Proc. of Second National Conference on Software Engineering, Berkeley, California (October 1976).

M.L. Shooman, "Need and Use of Analytical Methods in Reliability," Proc. of Joint Automatic Control Conference (IFAC), San Francisco, California (June 1977).

M.L. Shooman, "The Spectre of Software Reliability and Its Exorcism," Proc. of Joint Automatic Control Conference (IFAC), San Francisco, California (June 1977).

M.L. Shooman and H. Ruston, "Cost Reducing, High Reliability Programming Techniques," Proc. of 1976 ORSA/TIMS Joint National Meeting, Miami, Florida (November 1976).

M.L. Shooman and S. Sinkar, "Generation of Reliability and Safety Data by Analysis of Expert Opinion," Proc. of 1977 Annual Reliability and Maintainability Symposium, Philadelphia, Pennsylvania (January 1977).

T.S. Sundresh, F.A. Cassara and H. Schachter, "Suppression of Interchannel Interference in FM Receivers," Proc. of IEEE Conference on Communications and Power, Montreal, Canada, 63-65 (October 1976).

T. Tamir and S.T. Peng, "Network Methods for Integrated Optics Devices," Proc. of International Conference on Application of Holography and Optical Data Processing, Jerusalem, Israel (Oxford: Pergamon Press, 1977).

W.C. Wang, K.C. Whang and L.S. Rosenheck, "Optical Imaging and Memory Using a Si Diode Storage Correlator," Proc. of 1976 IEEE Ultrasonics Symposium, Annapolis, Maryland, 352-356 (October 1976).

W.C. Wang, H. Schachter, F.A. Cassara, K.C. Whang and L.S. Rosenheck, "Information Storage Using Diode Arrays SAW Correlator," Extended Abstracts of Electrochemical Society 151st Meeting, 77-1, 115-117 (May 1977).

T.Q. Yip, S.P. Kuo and B.R. Cheo, "Experimental Study of Time Evolution of Parametrically Excited Instabilities in a Magneto Plasma," Proc. of IEEE International Conference on Plasma Science, Troy, New York (May 1977).

TECHNICAL REPORTS

H. L. Bertoni, "Applications of Beam Shift Phenomena to Non-destructive Testing" (POLY EE/EP 76-015).

E.S. Cassedy, "A Theoretical Study of Injection Tuning of Optical Parametric Oscillators," Final Report to Los Alamos Scientific Lab. (LASL Ref. No. NP6-21964).

Z.W. Chang, L.B. Felsen and A. Hessel, "Surface Ray Methods for Mutual Coupling in Conformal Arrays on Cylindrical and Conical Surfaces" (POLY EE/EP 76-012).

B.R. Cheo and Z. Zabar, "OH System and Operation Concerning the Tokamak Fusion Test Reactor," Final Report for EBASCO at Princeton University (No. EP-M-1105, 4C1).

B.R. Cheo and Z. Zabar, "Using Diode Rectifiers in the TF Power Supplies Concerning the TFTR," Final Report for EBASCO at Princeton University (No. EP-M-1678, 4D3).

Y. Fujimoto and E.A. Mishkin, "Analysis of Spherically Imploding Shocks" (POLY EE/EP 77-026).

A. Hessel, S. Choudhary, L.B. Felsen and S.Y. Shin, "The Analysis of Propagation on Periodically Slotted Coaxial Cable Above Lossy Ground" (POLY EE/EP 77-025).

C.L. Hsu, L. Shaw and S. Tyan, "Reliability Applications of Multivariate Exponential Distributions" (POLY EE 77-036).

S.P. Kuo and B.R. Cheo, "Studies of Parametric Decay Instabilities in Magneto Plasmas" (POLY EE/EP 77-027).

A.E. Laemmel and M.L. Shooman, "Statistical (Natural) Language Theory and Computer Program Complexity" (POLY EE/EP 76-020).

A.E. Laemmel, "Background Memory Circuits" (POLY EE/EP 77-024).

T.T. Lee and E.J. Smith, "On the Partitioning Representation for Finite Sets and Its Application" (POLY EE/EP 76-009).

W.R. McShane, K.W. Crowley and L. Shaw, "Management of Vehicular Traffic Facilities for Better Transit Movement," Department of Transportation Planning and Engineering, Polytechn. Inst. of New York (VMTA-76-2-1).

E.A. Mishkin, "Radiation from a Small Sphere" (POLY EE/EP 76-014).

E.A. Mishkin, "Radiation Field of a Small Laser Shocks Compressed Sphere" (POLY EE/EP 76-017).

E.A. Mishkin and Y. Fujimoto, "Analysis of a Cylindrical Imploding Shock Wave" (POLY EE/EP 77-029).

E.A. Mishkin and Y. Fujimoto, "Density Saturation Induced by Spherical Implosion" (POLY EE/EP 77-030).

E.A. Mishkin and Y. Fujimoto, "Energy and Collapse Time Considerations," Spherical Imploding Shock (POLY EE/EP 77-032).

B. Rudner, "Seeding-Tagging Estimation of Software Errors; Models and Estimates" (POLY EE/EP 76-019).

L. Shaw, "Measurement and Control of Queues and Other Markov Systems" (POLY EE/EP 76-021).

M.L. Shooman and H. Ruston, "Summary of Technical Progress–Software Modeling Studies" (POLY EE/EP 76-013).

S. Sinkar and M.L. Shooman, "The Use of Consensus in Analytical Safety" (POLY EE/EP 76-018).

DISTRIBUTION LIST
Contract No. F44620-74-C-0056
Report on the Joint Services Electronics Program for the Period Ending 14 Sept. 1977

DEPARTMENT OF DEFENSE

Director
National Security Agency
ATTN: Dr. T. J. Beahn
Ft. George C. Meade, MD 20755

Defense Documentation Center (12)
ATTN: DDC-TCA (Mrs. V. Caponio)
Cameron Station
Alexandria, VA 22314

Asst. Dir., Electronics & Computer Sciences
Ofc. DDR & E
The Pentagon
Washington, DC 20315

DEPARTMENT OF THE ARMY

Commandant
US Army Air Defense School
ATTN: ATSAD-T-CSM
Ft. Bliss, TX 79916

Commander
US Army Armament R&D Command
ATTN: DRSAR-RD
Dover, NJ 07801

Commander
US Army Ballistics Research Laboratory
ATTN: DRXRD-BAD
Aberdeen Proving Ground
Aberdeen, MD 21005

Commandant
US Army Command and Gen. Staff Coll.
ATTN: Acquisitions, Lib. Div.
Ft. Leavenworth, KS 66027

Commander
US Army Communication Command
ATTN: CC-OPS-PD
Ft. Huachuca, AZ 85613

Commander
US Army Materials and Mechanics Research Center
ATTN: Chief, Materials Sciences Div.
Watertown, MA 02172

Commander
US Army Material Development & Readiness Command
ATTN: Technical Library, Rm. 7S 35
5001 Eisenhower Avenue
Alexandria, VA 22333

Commander
US Army Missile R&D Command
ATTN: Chief, Document Section
Redstone Arsenal, AL 35809

Commander
US Army Satellite Comm. Agy.
Ft. Monmouth, NJ 07703

Commander
US Army Security Agy.
ATTN: IARD-T
Arlington Hall Station
Arlington, VA 22212

Project Mgr.
ARTADS
EAI Building
West Long Branch, NJ 07764

Commander Director
Atmospheric Sciences Lab. (ECOM)
ATTN: DRSEL-BL-DD
White Sands Missile Range, NM 88002

Commander
USAEC
ATTN: DRSEL-NL-C (Dr. H.S. Bennett)
Ft. Monmouth, NJ 07703

Director
TRI-TAC
ATTN: TT-AD (Mrs. Briller)
Ft. Monmouth, NJ 07703

Commander
USAEC
ATTN: DRSEL-CT-L (Dr. R. Buser)
Ft. Monmouth, NJ 07703

Director
Electronic Warfare Lab. (ECOM)
ATTN: DRSEL-WL-MY
White Sands Missile Range, NM 88002

Executive Secretary, TAC JSEP
US Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709

Commander
Frankford Arsenal
Deputy Dir., Pitman-Dunn Lab.
Philadelphia, PA 19137

Project Manager
Ballistic Missile Defense Program Ofc.
ATTN: DACS-DMP (Mr. A. Cold)
1300 Wilson Blvd.
Arlington, VA 22209

Commander
Harry Diamond Laboratories
ATTN: Mr. John E. Rosenberg
2800 Powder Mill Road
Adelphi, MD 20783

HQDA (DAMA-ARZ-A)
Washington, DC 20310

Commander
USAEC
ATTN: DRSEL-TL-E (Dr. J.A. Kohn)
Ft. Monmouth, NJ 07703

Commander
USAEC
ATTN: DRSEL-TL-EN
(Dr. S. Kroenenberg)
Ft. Monmouth, NJ 07703

Commander
USAEC
ATTN: DRSEL-NL-T (Mr. R. Kulinyi)
Ft. Monmouth, NJ 07703

Commander
USAEC
ATTN: DRSEL-NL-B (Dr. E. Lieblein)
Ft. Monmouth, NJ 07703

Commander
USAEC
ATTN: DRSEL-TL-MM (Mr. N. Lipetz)
Ft. Monmouth, NJ 07703

Commander
USAEC
ATTN: DRSEL-RD-O
(Dr. W. S. McAfee)
Ft. Monmouth, NJ 07703

Director
Night Vision Lab.
ATTN: DRSEL-NV-D
Ft. Belvoir, VA 22060

COL Robert Noce
Sr. Standardization Rep.
US Army Standardization Group, Canada
Canadian Force Headquarters
Ottawa, Ontario
CANADA K1A K2

Commander
USAEC
ATTN: DRSEL-NL-B (Dr. D.C. Pearce)
Ft. Monmouth, NJ 07703

Commander
Picatinny Arsenal
ATTN: SMUPA-TS-T-S
Dover, NJ 07801

Dr. Sidney Ross
Technical Director
SARFA-TD
Frankford Arsenal
Philadelphia, PA 19137

Commander
USAEC
ATTN: DRSEL-NL-RH-1 (Dr. F. Schwering)
Ft. Monmouth, NJ 07703

Commander
USAEC
ATTN: DRSEL-TL-I (Dr. C.C. Thornton)
Ft. Monmouth, NJ 07703

US Army Research Office (3)
ATTN: Library
P.O. Box 12211
Research Triangle Park, NC 27709

Director
Div. of Neuropsychiatry
Walter Reed Army Institute of Research
Washington, DC 20012

Commander
White Sands Missile Range
ATTN: STEWS-ID-R
White Sands Missile Range, NM 88002

DEPARTMENT OF THE AIR FORCE

Mr. Robert Barrett
RADC/ETS
Hanscom AFB, MA 01731

Dr. Carl E. Baum
AFWL (ES)
Kirtland AFB, NM 87117

Dr. E. Champagne
AFAL/DH
Wright-Patterson AFB, OH 45433

Dr. R.P. Dolan
RADC/ETSD
Hanscom AFB, MA 01731

Mr. W. Edwards
AFAL/TE
Wright-Patterson AFB, OH 45433

Professor R.E. Fontana
Head Dept. of Electrical Eng.
AFIT/ENE
Wright-Patterson AFB, OH 45433

Dr. Alan Corscadden
AFAPL/POD
Wright-Patterson AFB, OH 45433

USAF European Ofc. of Aerospace Res
ATTN: Major J. Gorrell
Box 14, FPO, NY 09510

LTC Richard J. Gowen
Dept. of Electrical Engineering
USAF Academy, CO 80840

Mr. Murray Kesselman (ISCA)
Rome Air Development Center
Griffiss AFB, NY 13441

Dr. C. Knausenberger
Air Force Member, TAC
Air Force Ofc. of Scientific Research
(AFSC) AFOSR/NE
Bolling Air Force Base, DC 20332

Dr. L. Kravitz
Air Force Member, TAC
AFOSR
(AFSC) AFOSR/NE
Bolling Air Force Base, DC 20332

Mr. R.D. Larson
AFAL/DHR
Wright-Patterson AFB, OH 45433

Dr. Richard B. Mack
RADC/ETER
Hanscom AFB, MA 01731

Mr. John Mottsmith (MCIT)
HQ ESD (AFSC)
Hanscom AFB, MA 01731

Dr. Richard Picard
RADC/ETSL
Hanscom AFB, MA 01731

Dr. J. Ryles
Chief Scientist
AFAL/CA
Wright-Patterson AFB, OH 45433

Dr. Allan Schell
RADC/ETE
Hanscom AFB, MA 01731

Mr. H.E. Webb, Jr. (ISCP)
Rome Air Development Center
Griffiss AFB, NY 13441

LTC G. Wepfer
AFOSR
(AFSC) AFOSR/NP
Bolling Air Force Base, DC 20332

LTC G. McKemie
AFOSR
(AFSC) AFOSR/NM
Bolling Air Force Base, DC 20332

DEPARTMENT OF THE NAVY

Dr. R.S. Allgaier
Naval Surface Weapons Center
Code WR-33
White Oak
Silver Spring, MD 20910

Naval Weapons Center
ATTN: Code 5515 - H.F. Blazek
China Lake, CA 93555

NOTE: One (1) copy to each addressee unless otherwise indicated

Dr. H.L. Blood
Technical Director
Naval Undersea Center
San Diego, CA 95152

Naval Research Laboratory
ATTN: Code 5200 - A. Brodzinsky
4555 Overlook Avenue, SW
Washington, DV 20375

Naval Research Laboratory
ATTN: Code 7701 - J.D. Brown
4555 Overlook Avenue, SW
Washington, DC 20375

Naval Research Laboratory
ATTN: Code 5210 - J.E. Dove
4555 Overlook Avenue, SW
Washington, DC 20375

Naval Research Laboratory
ATTN: Code 5410 - J.R. Davis
4555 Overlook Avenue, SW
Washington, DC 20375

Naval Ocean Systems Center
ATTN: Code 70 - W.J. Deike
271 Catalina Boulevard
San Diego, CA 92152

Naval Weapons Center
ATTN: Code 6011 - F.C. Essig
China Lake, CA 93555

Naval Research Laboratory
ATTN: Code 5510 - W.L. Faust
4555 Overlook Avenue, SW
Washington, DC 20375

Naval Research Laboratory
ATTN: Code 2627 - Mrs. D. Folen
4555 Overlook Avenue, SW
Washington, DC 20375

Dr. Robert R. Fossum
Dean of Research
Naval Postgraduate School
Monterey, CA 93940

Dr. C. G. Gould
Technical Director
Naval Coastal System Laboratory
Panama City, FL 32401

Naval Ocean Systems Center
ATTN: Code 7203 - V.E. Hildebrand
271 Catalina Boulevard
San Diego, CA 92152

Naval Ocean Systems Center
ATTN: Code 753 - P.H. Johnson
271 Catalina Boulevard
San Diego, CA 92152

Donald E. Kirk
Professor and Chairman
Electronic Engineer, SP-304
Naval Postgraduate School
Monterey, CA 93940

Naval Air Development Center
ATTN: Code 01 - Dr. R.K. Labb
Johnsville
Warminster, PA 18974

Naval Research Laboratory
ATTN: Code 5270 - B.D. McCombe
4555 Overlook Avenue, SW
Washington, DC 20375

Capt. R.B. Meeks
Naval Sea Systems Command
NC #3
2531 Jefferson Davis Hwy
Arlington, VA 20362

Dr. H.J. Mueller
Naval Air Systems Command
Code 310
JP #1
1411 Jefferson Davis Hwy
Arlington, VA 20360

Dr. J.H. Mills, Jr.
Naval Surface Weapons Center
Electronics Systems Dept.
Code DF
Dahlgren, VA 22448

Naval Ocean Systems Center
ATTN: Code 702 - H.T. Mortimer
271 Catalina Boulevard
San Diego, CA 92152

Naval Air Development Center
ATTN: Technical Library
Johnsville
Warminster, PA 18974

Naval Ocean Systems Center
ATTN: Technical Library
271 Catalina Boulevard
San Diego, CA 92152

Naval Research Laboratory
Underwater Sound Reference Div.
Technical Libra
P.O. Box 8337
Orlando, FL 32806

Naval Surface Weapons Center
ATTN: Technical Libra y
Code DX-21
Dahlgren, VA 22448

Naval Surface Weapons Center
ATTN: Technical Libra
Building 1-330, Code WX-40
White Oak
Silver Spring, MD 20910

Naval Training Equipment Center
ATTN: Technical Library
Orlando, FL 32813

Naval Undersea Center
ATTN: Technical Library
San Diego, CA 92152

Naval Underwater Systems Center
ATTN: Technical Library
Newport, RI 02840

Office of Naval Research
Electronic & Solid State Sciences
Program (Code 427)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Mathematics Program (Code 432)
900 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Naval Systems Division
Code 220 221
300 North Quincy Street
Arlington, VA 22217

Director
Office of Naval Research
New York Area Office
715 Broadway, 5th Fl.
New York, NY 10003

Office of Naval Research
San Francisco Area Office
760 Market Street, Rm. 447
San Francisco, CA 94102

Director
Office of Naval Research Branch Ofc.
495 Summer Street
Boston, MA 02210

Director
Office of Naval Research Branch Ofc.
536 South Clark Street
Chicago, IL 60605

Director
Office of Naval Research Branch Ofc.
1030 East Green Street
Pasadena, CA 91101

Mr. H. R. Riedl
Naval Surface Weapons Center
Code WR-34
White Oak Laboratory
Silver Spring, MD 20910

Naval Air Development Center
ATTN: Code 202 - T.J. Shopple
Johnsville
Warminster, PA 18974

Naval Research Laboratory
ATTN: Code 5403 - J.E. Shore
4555 Overlook Avenue, SW
Washington, DC 20375

A.L. Slafkovsky
Scientific Advisor
Headquarters Marine Corps
MC-RD-1
Arlington Annex
Washington, DC 20380

Harris B. Stone
Office of Research, Development, Test
and Evaluation
NOP-937
The Pentagon, Rm. 5D760
Washington, DC 20350

Mr. L. Sumney
Naval Electronics Systems Command
Code 3042, NC #1
2511 Jefferson Davis Hwy
Arlington, VA 20360

David W. Taylor
Naval Ship Research & Development Ctr.
Code 522.1
Bethesda, MD 20084

Naval Research Laboratory
ATTN: Code 4105 - Dr. S. Teitler
4555 Overlook Avenue, SW
Washington, D.C. 20375

Lt. Cdr. John Turner
NAVMAT 0343
CP #5, Rm. 1044
2211 Jefferson Davis Hwy
Arlington, VA 20360

Naval Ocean Systems Center
ATTN: Code 740 - H. H. Wieder
271 Catalina Boulevard
San Diego, CA 92152

Dr. W.A. Von Winkle
Associate Technical Director
for Technology
Naval Underwater Systems Center
New London, CT 06320

Dr. Cernot M. R. Winkler
Director, Time Service
US Naval Observatory
Massachusetts Avenue at
34 Street, NW
Washington, DC 20390

Officer in Charge
Carderock Lab
Code 13 - G.H. Gleissner
David Taylor Naval Ship R&D Center
Bethesda, MD 20084

Dr. Ray F. Potter
3767 Talbot Street
San Diego, CA 92106

Director, ONR
San Francisco Area Office
1 Hallidie Plaza
Suite 601
San Francisco, CA 94102

OTHER GOVERNMENT AGENCIES

Dr. Howard W. Etzel
Deputy Director
Division of Materials Research
National Science Foundation
1800 G Street
Washington, DC 20550

Mr. J.C. French
National Bureau of Standards
Electronics Technology Div.
Washington, DC 20234

Dr. Jay Harris
Program Director
Devices and Waves Program
National Science Foundation
1800 G Street
Washington, DC 20550

Los Alamos Scientific Laboratory
ATTN: Reports Library
P.O. Box 1663
Los Alamos, NM 87544

Dr. Dean Mitchell
Program Director, Solid-State Physics
Div. of Materials Research
National Science Foundation
1800 G Street
Washington, DC 20550

Mr. F.C. Schwenk, RD-T
NASA
Washington, DC 20547

M. Zane Thornton
Deputy Director Institute for
Computer Sciences & Technology
National Bureau of Standards
Washington, DC 20234

NON-GOVERNMENT AGENCIES

Director
Columbia Radiation Laboratory
Columbia University
538 West 120 Street
New York, NY 10027

Director
Coordinated Science Laboratory
University of Illinois
Urbana, IL 61801

Director of Laboratories
Division of Engineering &
Applied Physics
Harvard University
Pierce Hall
Cambridge, MA 02138

Director
Electronics Research Center
The University of Texas
Engineering-Science Bldg. 314
Austin, TX 78712

Director
Electronics Research Laboratory
University of California
Berkeley, CA 94720

Director
Electronics Sciences Laboratory
University of Southern California
Los Angeles, CA 90007

Director
Microwave Research Institute
Polytechnic Institute of New York
333 Jay Street
Brooklyn, NY 11201

Director
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139

Director
Stanford Electronics Laboratory
Stanford University
Stanford, CA 94305

Stanford Ginzton Laboratory
Stanford University
Stanford, CA 94305

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(9) Scientific interim rept.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION'S CATALOG NUMBER

15 Sep 76-14 Sep 77,

(6) 4. TITLE (and Subtitle)

PROGRESS REPORT No. 42 TO THE JOINT SERVICES TECHNICAL ADVISORY COMMITTEE.

5. TYPE OF REPORT & PERIOD COVERED

Scientific Interim

(14) 6. PERFORMING ORG. REPORT NUMBER

POLY-MRI-452-42-77

(10) 7. AUTHOR(s)

Arthur A. Oliner

(15) 8. CONTRACT OR GRANT NUMBER(s)

F 44620-74-C-0056

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Polytechnic Institute of New York
Microwave Research Institute
333 Jay Street, Brooklyn, NY 11201

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

Project Nos. (16) 4751
61102F
681305

(11) 11. CONTROLLING OFFICE NAME AND ADDRESS

Joint Services Electronics Program through
Dir. of Electronic and Solid State Sciences
Air Force Office of Scientific Research (NE)
Bolling AFB, Washington, DC 20332

12. REPORT DATE

Nov[redacted]77

13. NUMBER OF PAGES

614

(12) 610 p.

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)

UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

1. This document has been approved for public release and sale; its distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

TECH, OTHER

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Acoustic Correlators
Antennas
Communications
Computer Communications
Computers
Control Theory
Data Processing
Electric Power Engineering
Electromagnetics
Electronics
Electrophysics
Energy Conversion

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report is the forty-first in a series of progress reports to the Joint Services Technical Advisory Committee since inception of the Joint Services Electronics Program at the Polytechnic Institute of Brooklyn in July 1955. The report, now being issued annually, summarizes research accomplished under the aegis of the Microwave Research Institute, and reflects the impact of the Joint Services Electronics Program on the research activities of faculty and students of the Institute. The program covers a broad spectrum ranging from basic theoretical physics, mathematics, and engineering, to experimental

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

409 138

19.

Micro-Electronics
Microwave Acoustics
Network Theory
Optics
Quantum Electronics
Software Reliability and Engineering
Solid State and Materials
Systems Control
Waveguide Techniques
Wave-Matter Interactions

20.

investigations involving basic measurements, development of devices, and materials.

Each activity reports in summary fashion on specific results obtained during the report period, 15 September 1976 through 14 September 1977, with individual acknowledgement of the sponsorship which has contributed to the reported work. The report is organized into two major divisions. The first, Electrophysics, includes the topics of: Electromagnetics; Acoustics; Optics; Quantum Electronics; Solid State and Materials; Wave-Matter Interactions; and Electric Power Engineering. The second, Systems, includes the topics of: Communications; Computer and Computer-Communications Networks; Safety, Reliability and Software Engineering; Systems, Control and Networks; and Data Processing.

Topics representative of most of the Microwave Research Institute programs have been included in this report. The various sponsoring agencies and their contract numbers are individually acknowledged at the end of each item.

This comprehensive annual report to the Joint Services Technical Advisory Committee is made possible through the support of the Air Force Office of Scientific Research of the Air Force Systems Command; the Department of the Army, Army Research Office; and the Department of the Navy, Office of Naval Research; under Contract F44620-74-C-0056, Project No. 4751.

Reproduction in whole or in part is permitted for any purpose of the United States Government.

Qualified requestors may obtain additional copies from the Defense Documentation Center; all others should apply to the Clearinghouse for Federal Scientific and Technical Information.